Towards High-resolution Computational Approaches for Structure-based Drug Discovery

Jianing Li

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

© 2011
Jianing Li
All rights reserved

ABSTRACT

Towards High-resolution Computational Approaches for Structure-based Drug Discovery

Jianing Li

This dissertation describes new computational approaches at high resolution for practical structure-based drug discovery. It begins with a brief review of structure-based computational approaches for drug discovery in comparison with ligand-based ones, followed by a discussion of important applications in selecting drug-like compounds and predicting drug metabolites. Since three-dimensional target structures are crucial for structure-based drug discovery, a new methodology based on force fields for protein structure refinement was developed. This methodology employs the VSGB 2.0 energy model in combination with a robust protonation state assignment algorithm and efficient sampling strategies. High accuracy was obtained for predicting 2239 protein side chains and 115 14-20 residue loops. Given the precision and uniform robustness, this methodology is believed for the first time to be suitable to tackle practical problems in structure-based drug discovery.

The VSGB 2.0 energy model was then applied in the development of a new accurate approach (IDSite) for predicting P450-mediated drug metabolism, a problem of great practical interest for drug discovery. IDSite is able to efficiently model induced-fit effects using flexible docking and constrained refinements. Sites of metabolism are determined based on the physical interactions between a P450 enzyme and the ligand. Preliminary tests with 56 compounds displayed both low false positive and low false negative rates, which demonstrate the high potential of IDSite to be used in metabolism tests for drug discovery.

In conclusion, this dissertation presents new computational approaches at high resolution to problems related to structure-based drug discovery with unprecedented accuracy. Given such high accuracy, these approaches are very promising in addressing practical issues in pharmaceutical research and development, and in enhancing our capability in the search for new safe drugs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CYP	Cytochrome P450
ICDA	Interaction Cluster Decomposition Algorithm
LJ	Lennard-Jones
MC	Monte Carlo
MD	Molecular Dynamics
MM	Molecular Mechanics
OPLS	Optimized Potentials for Liquid Simulations
OPLS-AA	Optimized Potentials for Liquid Simulations-All Atom
PBE	Poisson-Boltzmann Equation
PDB	Protein Data Bank
PLOP	Protein Local Optimization Program
PPW	Protein Preparation Wizard
QM	Quantum Mechanics
QSAR	Quantitative Structure-Activity Relationship
RMSD	Root-Mean-Square Deviation
SGB	Surface Generalized Born
SOM	Sites of Metabolism
VDW	Van der Waals
VSGB 2.0	Optimized Variable Dielectric Surface Generalized Born version 2.0

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Professor Richard A. Friesner, for his excellent scientific guidance, constant support and continuous encouragement during my five-year graduate studies. His passion and creativity in research have greatly inspired me to pursue my independent research career.

I would like to thank members of my dissertation committee Professor Bruce J. Berne, Professor Barry Honig, Professor David Reichman, and Professor Ruben L. Gonzalez Jr., who have generously given their time and expertise to improve my work.

I would like to thank all coworkers from Schrödinger Inc., who have generously shared their knowledge and experience with me. Special thanks to Dr. Ramy Farid, Dr. Robert Abel, Dr. Kai Zhu, Dr. Yixiang Cao, Dr. Zhiyong Zhou, Dr. Tyler Day, Dr. Arteum Bochevarov, and Dr. David Rinaldo, who have worked with me together on research projects and from whom I learned to be a rigorous and creative researcher.

I would also like to thank all the previous and current members in the Friesner lab for helpful and inspiring discussions. Particular thanks to Dr. Li Tian, Dr. Suwen Zhao, Dr. Thomas Hughes, Michelle Hall, Joseph Bylund and Andrew Weisman. I am also grateful to Ms. Elizabeth Cusack, Ms. Alix Lamia, Ms. Danielle Farrell, and Mr. Calman Lobel, who provided a lot of support to my graduate studies.

Last but not the least, my gratitude goes to my family and friends. I would like to express special thanks for my parents, Guochang Li and Xiaozhi Huang. They have given me tremendous encouragement and understanding, which nothing could really substitute. I would also like to thank my fiancé, Severin T. Schneebeli, as a wonderful partner in research and in everyday life.

This dissertation is dedicated to

my Father, Mother, Fiancé, and all those,

whose support, encouragement, and personal sacrifice

have made this research possible.

Chapter 1. Introduction

Modern drug discovery is an extremely costly and lengthy process that involves many chemical and biological technologies. Among all these technologies, computational chemistry is becoming more and more widely used in identifying and optimizing molecules as potential drug candidates, helping to alleviate the considerable efforts required by the experiments. In the last decade, the advancement of high-performance computing (HPC) and the development of databases allowed medicinal chemists to access a much larger chemical space *in silico* than *in vitro*. This is leading to a gradual switch from traditional drug discovery approaches (such as semi-rational design, massive pharmacological screening, or even serendipity) towards more fruitful combinatorial approaches involving computer-aided techniques. Furthermore, while traditional non-virtual practices of drug discovery are facing problems such as high false positive rates,[4] soaring cost, and slow procedures, computational techniques are becoming very attractive complements. Computational drug discovery has the following major advantages over traditional experimental approaches: 1) improving the hit rate and potency of candidates,[4] 2) reducing the demanding usage of assays, apparatus, and labor, and 3) streamlining the drug discovery process. It is believed that technology of computational chemistry has already permeated today's pipelines of drug discovery, delivering new drug candidates at a faster pace and lower cost.[4-6]

The current practical computational approaches for drug discovery are generally categorized into (but are not limited to) ligand- and structure-based.[2,7,8] Ligand-based approaches rely on prior knowledge about what binds to the target. Such methods include pharmacophore recognition[9,10] and quantitative structure-activity relationships[11,12], which select the drug-like compounds according to the spatial arrangements and the descriptors based on the compound's chemical structure. So far, many ligand-based approaches have been successfully implemented

in a number of software packages (e.g. PHASE[13] and DOGS[14]), but the limitations associated with the field are obvious: dependence on known binding modes, sensitivity to the alignment with known binding compounds, and poor predictions of binding affinity.[2] In addition, it is rare that a compound obtained from an existing library (e.g. a hit or lead compound) becomes a drug without any modifications.[1] In most cases, chemical alterations to various extents (lead optimization) are needed in order to improve the binding affinity and the bioavailability. In the absence of three-dimensional target structures, such modifications can hardly be designed precisely, so that extensive synthesis and experimental tests are still needed to confirm the discovery.

In contrast to ligand-based approaches, structure-based ones consider the interactions between the small molecules (ligand) and the therapeutic target (receptor) explicitly, and determine the selection based on calculated affinity or selectivity.[15-17] The most popular structure-based computational approach is molecular docking, which finds compounds that fit into the binding site with desired chemical properties and has been implemented in many software packages (e.g. Glide[18,19] and Autodock[20]). The major input needed for structure-based drug discovery is the three-dimensional structure of the biological target, while the structures of the drug candidates can simply be obtained from databases or constructed *de novo*. On one hand, independent of any known binding molecule or binding mode, a structure-based approach allows a more extensive search in the vast chemical space and has a potential of leading to novel candidates.[5] On the other hand, structure-based approaches can also utilize the information from the known binding modes in order to narrow down the search and reduce the computational cost.[21] Further, structure-based modeling is able to provide predictions of new binding modes, binding affinity, metabolites, etc.,[6,22] which are important considerations for further optimization

of the compounds. During recent years, there has been a growing effort to apply structure-based computational approaches to practical drug discovery. The development of such approaches highly benefits from the advances in functional genomics and proteomics as well as the increasing number of structures in the Protein Data Bank (PDB). For those targets without experimentally available structures, comparative modeling is applicable to building the atomic resolution model of the target from known templates of homologous proteins.[23-25] The structure-based computational approaches have already contributed to more than 50 new compounds, which entered into clinical trials.[1] One of the major successful stories is the identification of HIV-1 protease inhibitors(Figure 1.1).[26,27]

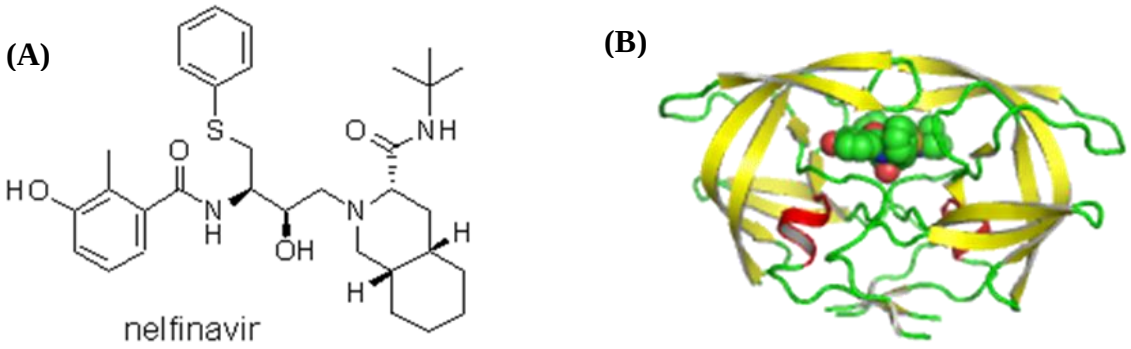


Figure 1.1. An example of an HIV-1 protease inhibitor, whose development benefited from computational structure-based techniques. **(A)** HIV-1 protease inhibitor nelfinavir. **(B)** co-crystallized structure of HIV-1 protease with nelfinavir. (PDBID: 3EKX).

It is worth pointing out that the ligand- and structure-based approaches are often used in combination to solve practical problems. Synergies can be achieved easily by combining the strength of both approaches: the ligand-based approach can be applied to prioritize or filter compounds for further structure-based investigation; the structure-based approach can also utilize the pharmacophore constraints for directed sampling. Since 2000 there have been already

plenty of published examples, such as the development of inhibitors of HIV-1 protease[26] and JNK3kinase.[8]

Recently, there has been a growing interest in employing high-resolution target structures to accurately model ligand-receptor interactions, because the precise information about interactions such as solvation effects, hydrogen bonding, π-π interactions, and cation-π interactions are important in deciding the ligand orientations as well as the energetic and conformational changes during binding. Prior work includes identifying structural water and small molecule binding sites,[28] computing binding affinity,[22] predicting conformational changes upon binding,[29] revealing conformational coupling in allosteric proteins,[30] and studying interactions between proteins and other macromolecules.[31,32] Based on the high-resolution target structures, the approaches are expected to accurately predict the binding mode and the binding affinity, which in turn requires leveraging the detailed information encoded in the structures and to accurately model the various energetic and conformational impacts. Ideally, such approaches are also expected to result in lower false positive rates (even lower false negative rates as well) as their resolution is high enough to distinguish subtle energetic and conformational changes. However, in spite of the recent success of computational structure-based drug discovery mentioned above, it is still very challenging to achieve accuracy sufficient to solve practical problems, mainly due to the scarcity of high-resolution target structures and the lack of accurate, efficient computational approaches at high resolution. Although modeling at Quantum Mechanics (QM) level is able to provide high accuracy, its computational cost is often overwhelming for macromolecules, especially due to the large number of conformations associated with each structure. Thus with regard to the current computational capacity, it is not a practical solution to apply QM methods for drug discovery tasks such as large scale virtual

screening. As an alternative, force field methods at the all-atom level are able to consider many structures at a relatively low cost, but the accuracy highly varies from case to case. This calls for the development of robust force field approaches with high accuracy for structure-based drug discovery.

Focusing on high-resolution modeling with force field methods, this dissertation describes research studies that develop accurate computational approaches for structure-based drug discovery. It covers the topics of target structure refinement and drug metabolite predictions, aiming to solve practical problems with high accuracy and efficiency. Since proteins represent around 90% of the current therapeutic targets, the focus of this dissertation is on proteins (as the typical targets), but we believe that our methodologies described herein are general, such that they can also be applied to other biomacromolecules.

The principle goal of structure-based drug discovery is to identify compounds with high binding affinities with the therapeutic target. As mentioned above, structures at atomic (or even sub-atomic) resolution are required in order to achieve this goal.[22] However, most currently known macromolecular structures do not fulfill these high resolution requirements. While the majority (~90%) of the experimental structures in the PDB (www.rcsb.org) have resolutions poorer than 2.0 Å, many of the current comparative models do not even reach an accuracy of 2.0 Å RMSD from the native structures, especially at low sequence identity (e.g. <30%).[33] Therefore, it is necessary to improve the low-resolution target models before they are used for tasks in drug discovery (such as virtual screening). Progressive computational refinement of the low resolution target structures represents a cost-effective and general solution to this problem, provided that accurate energy models and efficient sampling algorithms are available. To address these problems, Chapter 2 of this dissertation describes our new accurate VSGB2.0 energy

model and efficient sampling algorithms for protein structure refinement. The VSGB 2.0 energy model includes specific terms to consider solvation effects, hydrogen bonding, π-π packing, and self-contacts, while our sampling algorithms employ a novel dipeptide library to improve the efficiency. Our new methodology has been rigorously tested on large data sets of protein single side chain (>2,000) and loop predictions (>100) and shows great potential to provide high-resolution protein structures for structure-based drug discovery.

Apart from identifying drug-like compounds, it is also of great importance to predict the absorption, distribution, metabolism, excretion, and toxicity properties (ADMET) of these compounds.[6,34,35] Accurately predicting the ADMET of drug candidates computationally helps to eliminate compounds with undesirable properties such as poor absorption and toxic side effects and therefore reduces the risks to fail in clinical trials. In particular, predicting the metabolites of drug candidates is crucial in improving the pharmacokinetics and in avoiding the toxicity associated with metabolites. It is believed that 12 enzymatic systems are involved in drug metabolism, among which the P450 system is the major one responsible for phase I biotransformation.[36] So far, a number of P450 isoforms have been crystallized and their structures have been determined.[37] These experimental structures provide an excellent starting point for developing structure-based approaches to predict the P450-mediated drug metabolism. Therefore, it is not surprising that predicting P450 metabolism has become a rapidly growing field.[38-40] However, very few published methods and models can achieve both low false positive and low false negative rates in their predictions, and the inadequate accuracy of current methods has become a big challenge for practical use.[41] Herein, a highly accurate computational approach, IDSite, based on the interactions of the substrates and P450 enzymes is developed in Chapter 3. IDSite was developed on the basis of the VSGB 2.0 energy model and a smart directed sampling

protocol, which involves docking and refinement. The low false positive and false negative rates shown by tests of 56 compounds indicate the great potential of IDSite for practical drug discovery.

In summary, all of the methodology presented in this dissertation is likely to enhance and streamline the process of lead generation and optimization, and provide *in silico* solutions to discover better drugs with high activity and good safety profiles. Further, the methodology can also be used to construct valuable tools to investigate the mechanism of action[42] during target validation, which will further accelerate the drug discovery process.

Chapter 2. Protein Side Chain and Loop Predictions

2.1. Introduction

Knowledge of protein structure at atomic resolution is essential for modeling biological function and structure-based drug discovery approaches.[43-45] While the generation of experimental structures, propelled by high throughput crystallography,[46] continues to advance exponentially, the number of known protein sequences is growing even more rapidly.[47] Furthermore, for any given sequence, there may be a significant number of biologically relevant conformations, not to mention possible structural reorganization associated with ligand binding or with protein-protein interactions. Hence, it is unlikely that the entire universe of biologically relevant protein structural data can be accessed by exclusively experimental means.

Computational modeling represents the logical approach to constructing protein structures that are not experimentally available. The coverage of protein families continues to increase rapidly in PDB, which implies that the vast majority of protein structure prediction problems involve perturbation of a known structure by a relatively small RMSD. Homology modeling, using sequence and profile-based approaches,[25,48,491-3] continues to make great progress, and models with the correct architecture and low RMSD can be built for substantial fraction of interesting cases, particularly for pharmaceutically relevant targets where substantial experimental work on the protein family to which the target belongs (e.g. kinases) has typically been performed.[50,51]

However, to predict relative protein conformation energetics and protein-ligand binding affinities, very high resolution structures are required, and current homology models are often not quite good enough for this purpose (although the suitability varies depending upon the target and the specific project for which the structure will be employed).[52] The technology that would

address this problem is refinement of homology models, in which the RMSD of the homology model is progressively reduced until it is suitably close to the native structure. Such a refinement strategy in turn requires a sufficiently accurate potential energy function, including modeling of solvation effects and detailed physical chemistry of protein interactions (e.g. hydrogen bonding). If the potential energy surface has a free energy minimum that is distinct from the native structure, this implies a fundamental limitation on the RMSD that can be achieved. With an accurate potential surface, one is then left with the problem of sufficiently robust and comprehensive sampling of phase space, a challenging task given that the homology model may deviate from the native structure at any location in the protein.

There are various possible strategies that can be employed to carry out refinement. The most straightforward approach would be to perform a molecular dynamics simulation[53] using an all-atom protein model and explicit representation of aqueous solvent.[54] However, such simulations are very expensive computationally, and even assuming that the potential functions used in the simulation (which at present generally do not incorporate polarizability, for example) are adequate to yield the native structure as a free energy minimum, the effort required would be extremely large even for a small protein, and prohibitive for larger proteins and protein assemblies which constitute the great majority of biologically interesting systems.

An alternative approach to refinement is to utilize a continuum representation of solvent, along with an all-atom protein force field. Continuum approaches have two major advantages. First, it is not necessary to average over the positions of explicit water molecules, which generally requires very lengthy convergence times.[55] Secondly, conformational search methods, as opposed to molecular dynamics, can conveniently be employed in conjunction with a continuum solvation model.[56] Such methods can be many orders of magnitude more efficient

than molecular dynamics for locating the global free energy minimum of the model, because much larger steps in phase space can be taken. The generalized Born continuum solvent model,[57,58] in particular, is relatively inexpensive to evaluate, and is amenable to calculation of gradients, which are necessary for minimizations.

While considerable progress has been made, in our group and others, in advancing the state of the art in continuum solvation calculations,[59-62] two principal problems still remain. First, as in any energy model applied to high resolution protein structure refinement, the model must be accurate enough to actually improve homology models beyond their current level of resolution. In considering the accuracy of a force field plus a continuum solvation model, what matters is the potential energy surface defined by the model as a whole, as opposed to the individual components. The problem is in some ways more challenging than that of constructing force fields for explicit solvent simulation, because development of an accurate continuum model requires guessing and experimentation, since the functional form represents a reduction of the true Hamiltonian of the system to a non-rigorous model approximation (as is the case, for example, in density functional theory in electronic structure). Hence, the accuracy of continuum models requires continuous improvement, either by comparison with explicit solvent simulations, experimental data, or both. Secondly, the sampling problem remains one of great difficulty, although amenable to a wide range of algorithmic acceleration due to the ability to employ conformational search methods of various types.

In previous studies of PLOP, a series of improvements in both the energy model and sampling algorithms have been implemented in the program.[63-70] These improvements have enabled reasonable results to be obtained for loop predictions up to 20 residues. However, a non-trivial fraction of test cases continued to exhibit large RMSDs, in some cases accompanied by

large energy errors (defined as the energy gap between the predicted and minimized native structures). These results, while encouraging, reflected the fact that there was still important missing physics in the previous energy models.

In this chapter, a new energy model (VSGB 2.0) is described, which has been rigorously optimized by fitting to accurate experimental side chain and loop (11-13 residues) data, and contains a number of new terms not incorporated into the older functional form, as well as many re-optimized model parameters. The performance of the VSGB 2.0 model was evaluated by predicting structures for a set of 115 super long loops of 14-20 residues. At these lengths, alternative approaches in the literature uniformly display rapidly increasing RMSD errors,[71,72] a reflection of the greatly expanded conformational freedom associated with these very long loops; and in fact, there are very few prior studies in which loops of such lengths have been investigated systematically using a large data set. Remarkably, despite the exceptionally demanding test set (for which no parameter adjustment was made to improve agreement with experiment), a high degree of robustness, and small backbone and side chain RMSD, are demonstrated for 100% of the test cases. Achieving this level of accuracy requires other improvements besides the energy model, most prominently continued advances in the sampling algorithms, and application of a reliable approach to assigning protonation states, both of which are described in what follows. Given the precision and uniform robustness of the calculations, it is believed that the VSGB 2.0 model and sampling algorithms, for the first time, are suitable for successfully tackling the *real* problem defined above, refining homology models.

Very different philosophies have been used over the past decade to optimize and to evaluate atomic level protein models based on continuum solvation description. Alternatives have included fitting generalized Born (GB) models to Poisson-Boltzmann (PB) results (note that

the PB model itself has to be parameterized in some fashion),[73,74] exploring performance in the folding of small proteins,[75,76] and comparisons with explicit solvent simulations.[54,77] The present work is distinguished by the approach of fitting parameters to a large database of crystallographic single side chain and loop data (including a number of novel terms, one of which, the variable dielectric model, approximately incorporates polarization, and has proved to be extremely important in obtaining quantitatively useful results), and rigorously evaluating structural prediction accuracy for a large and demanding test set, the long loop data set described above. In our view, the use of these large training and test sets eliminates the possibility of overfitting, and provides confidence that the physics of the model is correct, and the right answers are being obtained for the right reason.

This chapter is organized as follows. The selection of the training (side chain prediction, loop prediction) and test (super long loop prediction) sets is first discussed. It is shown that the use of an improved training set (as compared to that employed in ref.[68]) turns out to be very important; the training set from previous work, despite the overall structural accuracy implied by the crystallography, contained side chains with ambiguous atom placement due to missing electron density in the crystallographic data. Then briefly review of algorithms for side chain and loop prediction is provided.[63,64,66] The VSGB 2.0 model is discussed in detail, as well as the optimization protocol based on single side chain and 11-13 residue loop predictions. Results for the test set of super long loop prediction (length of 14-20 residues) are then presented, along with the discussion of the results. Finally, in the conclusion, the results are summarized.

2.2. Methods and Materials

Selection of Data Sets

The rapid increase in the number of high resolution X-ray crystallographic structures has allowed us to build reliable data sets for training and testing the VSGB 2.0 model. In order to ensure the high quality, the data sets were selected based on the following criteria:

1. All structures are PDB X-ray crystallographic structures with low sequence identity (no more than 30% similarity) and high resolution (better than 1.0 Å for the side chain set and 2.0 Å for the loop sets).
2. Side chains or loops should not have alternative structures or missing heavy atoms.
3. Side chains or loops should not be affected by ligands. The distance between the side chain/loop and a ligand is defined as the shortest heavy atom distance. The minimum distance allowed to any organic ligands is 4.0 Å and to any metal ions is 6.5 Å.
4. The average B-factor should be lower than 35.00.
5. The real space R-factor (RSR) of each residue should be lower than 0.200.
6. All atoms of the side chains or loops should be found to occupy well defined peaks in the experimentally determined electron density when visualized.

With all these criteria, 2239 single side chains from 45 proteins for the side chain training set, 100 loops (length of 11-13 residues) from 72 proteins for the loop training set, and 115 super long loops (length of 14-20 residues) from 97 proteins for the loop test set have been collected.

Preparation of All-atom Models

As most crystallographic structures do not contain the hydrogen positions, it is necessary to add hydrogen atoms and determine the protonation states of ionizable residues for calculations at an atomic level of resolution. Additionally, the ambiguous orientation of Asn, Gln, and His

(due to the similar electron density of two alternative conformations rotated by 180°) also impairs the correct physics of the models. Therefore given the heavy-atom coordinates from X-ray crystallography, all-atom models for each protein were created using the Interaction Cluster Decomposition Algorithm (ICDA).[67] The ICDA assigns protonation states of ionizable residues, conformations of Asn, Gln, and His, and hydrogen positions of hydroxyls, by constructing clusters of potentially interacting side chains, enumerating a list of possible hydrogen bonding networks, and ranking these potential networks via energy evaluation. It is also worthy to mention that in this work the original ICDA algorithm reported in ref. [67] has been improved by using self-adjusted cluster sizes and more rigorous energy evaluation.

In addition to the protein, crystal environment, organic ligands, and metal ions were also taken into account for a fair comparison to the crystal structures obtained from experiments. Since the role of crystal environment and ligands in protein structure prediction has been extensively discussed,[78] their inclusion in our all-atom models was done in an automated fashion.

Single Side Chain Prediction Algorithm

Single Side Chain Prediction (SSCP) is defined as prediction of the conformation of one side chain with the rest of the protein fixed at the atomic positions of the native structure.[63,68] The algorithm exhaustively samples side chain conformations with a residue-specific rotamer library at a high resolution.[79] Clash-free conformations are evaluated and sorted according to the single point energy or the energy after minimization. The final prediction is determined by the lowest energy conformation, either with or without minimization. In this algorithm, all the conformations that remain after the steric clash check are kept for the evaluation stage. This is a modification to the original algorithm employed in previous publications, which prescreens all the candidates with a reduced energy score and clusters the remaining conformations to only

consider the cluster representatives. Since the total energy score is the only measure to evaluate all the conformations in the pool, the current version of SSCP algorithm is better able to provide a direct comparison of how well the energy models can distinguish the native from the non-native conformations in realistic applications such as loop prediction, where all of conformational space is considered, and minimization of the loop (which includes all side chain degrees of freedom) is employed. The optimized parameters obtained from SSCP fitting are discussed below.

Loop Prediction (11-13 residues) with Decoys

Our loop prediction (length of 11-13 residues) was carried out with decoys generated from the loop predictions described in previous work.[66] Each loop case contains thousands of conformations in the decoy set, representing a wide spread of samples in the conformational space. For the purpose of optimizing the energy model (more specifically the hydrophobic term), the minimized conformation with the lowest energy was selected as the prediction. Loop prediction of 11-13 residues was employed to optimize the hydrophobic term, because the hydrophobic term is much larger in a loop than in a side chain and thus more sensitive in loop prediction. The use of decoys avoids the costly sampling in a vast conformational space, greatly reduces the sampling cost, and consequently allows fast optimization of the hydrophobic term. The optimized functional form and parameters are discussed below.

Super Long Loop Prediction (14-20 residues) Algorithm

Our algorithm of Super Long Loop Prediction (SLLP) uses a hierarchical approach combined with an advanced sampling method to predict loops longer than 13 residues.[64,66,80] Compared to the shorter loop prediction, super long loop prediction requires more intensive sampling efforts. To improve both accuracy and efficiency, the algorithm uses an advanced sampling method based on a detailed dipeptide backbone rotamer library, which was first described in previous work with 17% improvement in sampling efficiency.[80]

In the SLLP algorithm, the loop candidates are first constructed without any constraint at the initial stage, while the rest of the protein remains the same as the native. An exhaustive search for possible loop conformations is carried out in one constrained refinement stages and a series of fixed stages. (Figure 2.1) At the end of each stage, the loop conformations generated in all the previously finished stages are ranked according to the energy after minimization, and the top ones without redundancy are sent to the next stage. Finally, the loop conformation with the lowest minimized energy is selected as the prediction.

Figure 2.1. Hierarchical sampling of super long loop predictions.

Each stage shown in Figure 2.1 contains multiple parallel single loop predictions, which build up the loop candidates, cluster them, and minimize and evaluate the centroid of each cluster. In this work, *cis*- rotamers are newly added to the dipeptide library originally described in ref. [80] which only contains *trans*- rotamers. Compared to the generic backbone library, the dipeptide

library depends on residue types and contains more information associated with the secondary structure. As different from many fragment assembly methods (e.g. the 3-residue and 9-residue fragments in Rosetta[81]), our dipeptide sampling is a unique method constructed from a large variety of high resolution protein crystallographic dataset. While a fragment in the fragment assembly methods are commonly used as a sliding window along a protein chain, our dipeptide sampling doesn't use the rotamers in an overlapping fashion during a single loop prediction. Besides, each dipeptide rotamer only has 5 backbone dihedral angles (phi and psi angles of the two residues and the omega angle between them) for backbone sampling, but the 3-residue and 9-residue fragments contains both backbone and side chain conformations for a quite different application in predicting small protein folding.

To accurately predict protein loops longer than 13 residues, it is extremely important to generate conformations as close as possible to the native at the early sampling stages. During the initial stage, a large number of loop structures are constructed from 5 parallel calculations with overlap factor thresholds 0.45, 0.50, 0.55, 0.60, and 0.65. (In previous work, 0.55, 0.60, 0.65, 0.70, and 0.75 were used.[80]) An overlap factor is defined as the ratio of distance between two atoms and the sum of their Van der Waals radii. Loop conformations are considered with clashes and then abandoned, if any pair of atoms is found with an overlap factor lower than the threshold. A high overlap factor threshold sometimes causes sampling failure in a confined space or generates similar loop candidates. Therefore, multiple low overlap factors in the initial stage are important to generate a wide variety of loop candidates for the later stages.

The fixed stages, which sample a sub-region of the loop, are crucial in our hierarchical method for super long loop prediction. During the fixed stages, a large conformational space is searched and the native-like conformations are enriched progressively among the top loop

candidates with the lowest energies. A "Fix N" stage (N=1, 2, 3…) only samples the residues that are outside of the fixed N residues in the loop. Since these N residues can start from either the C- or the N-terminus of the loop in question, all the N+1 combinations are considered in each "Fix N" stage. Sampling up to the "Fix 5" stage is adequate for the accurate prediction of many super long loops. However extending to the "Fix 10" stage successfully addressed several cases with serious sampling errors but also could lead to a relatively small number of cases with energy errors. Results of long loop predictions are presented with sampling up to "Fix 5" for most of the cases except several cases sampled up to "Fix 10". Identified by the further test, the cases which have possible energy errors with sampling up to "Fix 10" are then discussed in the discussion section.

Super Long Loop Prediction Method Incorporating Surrounding Side Chains

To further test the effectiveness of the VSGB 2.0 model, super long loop prediction incorporating surrounding side chains (SLLP-SS) was performed. The method has been previously described by Sellers *et al.*,[70] in which the surrounding side chains that have heavy atoms within 7.5 Å from any C_β atoms in the target loop are optimized simultaneously with the loop. In SLLP-SS, these surrounding side chains are temporarily removed when the backbone of the loop is being sampled. Then the side chains on the loop as well as those in the surroundings are put back and optimized by rotamer library sampling and minimization. Because the conformational phase space increases substantially when the surrounding side chains are optimized in addition to the loop, our most extensive hierarchical approach for long loop prediction is employed using the sampling stages "Fix 1" through "Fix 10".

The VSGB 2.0 Model

Energy Function

The VSGB 2.0 model provides a novel form of the energy function (Eq. 2.1) which contains the OPLS-AA protein force field[82] bonded and nonbonded terms, as well as a solvation term and a number of physics-based correction terms. The solvation free energy G_{sol} and the components of physics-based corrections $E_{corrections}$ are described in detail below.

$$G_{total} = \sum_{bonds} k_b (r - r_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2}[1 + \cos(n\phi - \delta)]$$
$$+ \sum_{impropers} k_\varphi (\varphi - \varphi_0)^2 + \sum_{electrostatics} \frac{q_i q_j}{r_{ij} \varepsilon_{in(ij)}} + \sum_{VDW} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right]$$
$$+ G_{sol} + \sum E_{corrections} \qquad \text{(Eq. 2.1)}$$

Solvent model

Solvation and desolvation effects are among the most important factors to determine a protein's global and local conformations in solvent.[83] The VSGB 2.0 model approximates the solvation free energy with an optimized implicit solvent model, which is based on Surface Generalized Born (SGB) model[57,58,60] and the variable dielectric (VD) treatment of polarization from protein side chains.[68] The SGB model, as an approximation to the Poisson-Boltzmann Equation (PBE), has been widely used in protein structure modeling. The variable dielectric treatment further improves the accuracy of the SGB model by varying the internal dielectric constants from 1.0 to 4.0 to incorporate the polarization effects. In this work the VD-SGB implicit solvent model, as an important component of the VSGB 2.0 model, has been further optimized via fitting to single side chain predictions.

A typical GB model, the solvation free energy (G_{sol}) is expressed as the sum of a cavity term (G_{cav}), a Van der Waals term (G_{vdw}) and a polarization term (G_{pol}) (Eq. 2.2).

$$G_{sol} = G_{cav} + G_{vdw} + G_{pol} \quad \text{(Eq. 2.2)}$$

The nonpolar solvent-solute interaction is usually represented by the sum of the cavity term and the Van der Waals term, which is considered proportional to the solvent-accessible surface area (SASA) (Eq. 2.3).

$$G_{cav} + G_{vdw} = \sigma \cdot SASA \quad \text{(Eq. 2.3)}$$

However, such a surface area model has been found insufficient to account for the nonpolar solvation effect (e.g. dispersion) in many previous studies. Therefore in our VSGB 2.0 model, the nonpolar contribution of solvation is calculated by a parameterized hydrophobic term described in detail below. The polar solvent-solute interaction is represented by the polarization term which depends on the solvent and internal dielectric constants, the partial charges, and f_{GB} (Eq. 2.4)

$$G_{pol} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{in(ij)}} - \frac{1}{\varepsilon_{sol}}\right)\sum_{i<j}\frac{q_i q_j}{f_{GB}} \quad \text{(Eq. 2.4)}$$

$$f_{GB} = \sqrt{r_{ij}^2 + \alpha_{ij}^2 e^{-D}} \quad \text{(Eq. 2.5)}$$

$$D = \frac{r_{ij}^2}{\left(2\alpha_{ij}\right)^2} \quad \text{(Eq. 2.6)}$$

f_{GB} is a function of the distances between two atoms (r_{ij}) and their generalized Born radii (α_i and α_j) of the form described in ref. [68]. The internal dielectric constant $\varepsilon_{in\ (ij)}$ can vary from 1.0 to 4.0 as the maximum value of the internal dielectric constants of atom *i* and atom *j*. The Born alpha for the screen term is calculated as $\alpha_{ij} = \sqrt{\alpha_i \alpha_j}$. Table 2.1 shows the assignment of internal dielectric constants in the optimized VD-SGB model and in the original VD-SGB model.

Table 2.1. Internal dielectric constants of the original and optimized VD-SGB model.[a]

Residue	Lys	Glu	Hip[b]	Asp	Arg	His[b]	Other
Original VD-SGB Model	4.00	3.00	3.00	2.00	2.00	2.00	1.00
Optimized VD-SGB Model	3.85	2.78	2.86	2.44	2.11	1.00	1.00

[a.] The assignment of internal dielectric constants is based on an atom-based scheme: only the charged atoms and the atoms adjacent to the charged ones are assigned with values greater than 1.00.
[b.] Hip: protonated histidine; His: neutral histidine.

The optimized VD-SGB model reduces the values of internal dielectric constants for Lys, Glu, protonated His and neutral His while increases the values for Asp and Arg. These changes, although derived from parameterization, actually incorporate a better physical picture from two main aspects: first, the dielectric constant assigned to the neutral His is adjusted to be identical to other non-charged amino acid residues; second, the internal dielectric constants for the acidic amino acid residues (Asp and Glu) are tuned to have closer values, more consistent with their chemical similarity.

Hydrogen Bonding Correction

An accurate description of hydrogen bonds is critical to predicting protein structure at high resolution. While a conventional fixed charge force field such as OPLS-AA does a reasonable job of getting the magnitude of hydrogen bonding interactions right,[84] it is limited by having an atomic point charge description of electrostatics, as opposed to a more accurate multipole or lone pair description. One approach would be to improve the electrostatics by explicit addition of such higher order terms, as has been done in the AMOEBA force field.[85] An alternative is to use an empirical functional form to enforce hydrogen bond angles and distances, fitting to experimental PDB data. We have chosen to use the latter approach, following work of Baker and coworkers[86] as in the spirit of exploiting the large amount of experimental structural data in the PDB to achieve the highest possible accuracy for protein structure specifically. The

new terms are added as a correction to the existing OPLS-AA force field and as part of the VSGB 2.0 model, and the parameters are optimized by fitting to improve single side chain prediction accuracy, as is discussed below.

The hydrogen bonding correction E_{HB} term is a function of distances, angles and atom types (Eq. 2.7). As the implicit solvent model is used, this correction is not applied to protein-solvent hydrogen bonding, which can cause underestimation of the interaction between protein and the first-shell solvent.

$$E_{HB} = \sum_i \sum_j \alpha_i \alpha_j \exp[-\left(r^{HA} - r_0\right)^2] \cos\left(\theta^{DHA} - \theta_0\right) \qquad \text{(Eq. 2.7)}$$

where *i* and *j* are heavy atoms involved in a hydrogen bond. The parameters of hydrogen bonding geometry r_0 (optimal distance between the hydrogen atom and the acceptor atom, 1.94Å) and θ_0 (optimal angle of the donor atom, the hydrogen atom and the acceptor atom, 160 °) were adopted from the Density Functional Theory (DFT) optimized formamide dimer and acetamide dimer.[87] α_i and α_j are coefficients related to the roles of the heavy atoms playing in a hydrogen bond, one as a donor while the other as an acceptor. For an atom *i*, α_i is assigned based on the following rules:

1) Positively charged nitrogen atoms in the side chains of Lys, Arg, charged His, and the N-terminal backbone are strong donors, α_i = 1.5.
2) Negatively charged oxygen atoms in the side chains of Asp, Glu, and the C-terminal backbone are strong acceptor, α_i = 1.5.
3) Polar atoms in the side chains of Ser, Thr, Asn, Gln, neutral His, Tyr, and Trp can be either weak donors or acceptors. The assignment is dependent on the paired atom *j*: if atom *j* is a strong donor, atom *i* is a weak acceptor; if atom *j* is a strong acceptor and atom *i* has at least one bonded hydrogen atom, atom *i* is a weak donor; otherwise, atom *i* is

counted twice as a weak donor and a weak acceptor while atom *j* as a weak acceptor and a weak donor respectively (as long as both have hydrogen atoms). For the weak donor, α_i = 0.5; for the weak acceptor, α_i = 0.5.

4) A neutral backbone oxygen atom is considered as a weak acceptor while a neutral backbone nitrogen atom is considered as a weak donor.

The hydrogen bonding correction depends on the distance between hydrogen atom and the acceptor atom r_{HA}, as well as the angles formed by donor atom, hydrogen atom and acceptor atom θ_{DHA} (Figure 2.2). Such a design involves the most relevant atoms in a hydrogen bond while being easy to implement and inexpensive to calculate.

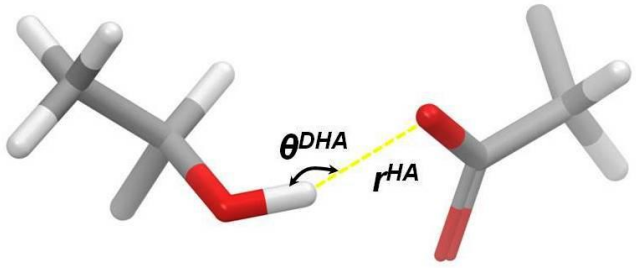


Figure 2.2. Geometry variables in hydrogen bonding correction.

π-π Packing Correction

π-π stacking is one of the main driving forces to stabilize vertical base stacks in DNA[88,89] and the hydrophobic cores in proteins.[90] It also plays important chemical and biological roles in processes like self-assembly[91] and molecular recognition.[92-94] In contrast to covalent bonds and hydrogen bonds, π-π interactions are slightly directional with diverse preferred configurations, such as parallel stacking (sandwich and parallel-displace) and T-stacking (also known as T-shape),[93,95] which are generally referred to as π-π packing interactions in this dissertation.

However, it is difficult to rigorously separate packing interactions from other nonbonded terms such as Van der Waals interactions. Two approaches can be used to provide estimates of

the magnitude of the packing effects; observation of π-π interactions in native protein structures, via crystallography, and quantum chemical calculations of such interactions, using high level theories that adequately capture electron correlation effects. For example, Burly and Petsko found that the preferential distance for π-π packing interaction is 4.5~7.0 Å and the preferential dihedral angles are close to 90 ° in 34 protein crystallographic structures.[96] An estimation of the free energy contribution from Burly and Petsko is between 0.6~1.3 kcal/mol. The quantum level study of Jurecka *et al.* gave a higher estimation as 2.5~7.0 kcal/mol for aromatic amino acid dimers.[97]

Our own investigations involve an analysis of our side chain and loop data sets, and comparison of predicted and native structures using our previous energy model, which did not contain an explicit π-π packing term. When such a model is used, the native structures contain a systematically higher percentage of π-π interactions (estimated by geometrical criteria) than is seen in predicted structures. These observations have motivated the development of a stacking term, empirically optimized to reproduce side chain structures, and then tested via loop prediction.

Given the insufficient treatment of π-π packing interaction in standard force field methods,[84] it would be useful to design a new π-π packing correction to improve the accuracy of the previous energy model. Here, we present an explicit π-π packing correction in the VSGB 2.0 model for pairs of amino acid side chains including the conventional aromatic ones such as Phe, Tyr, His and Trp as well as the Y-aromatic structures such as Arg,[92] Asn and Gln. To reduce the complexity of the algorithm, only side chain-side chain packings are considered, although π-π stacks also exist in interactions involving protein backbone.

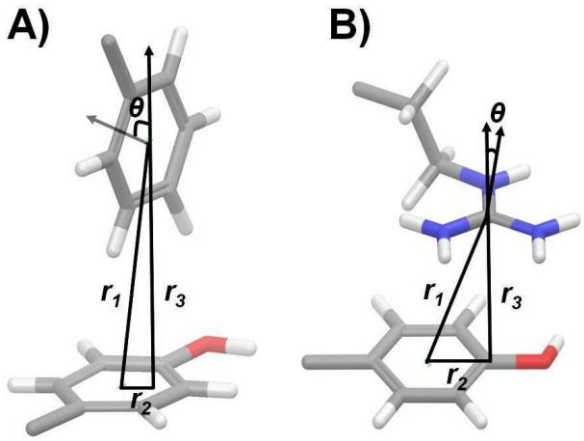


Figure 2.3. Geometry variables in π-π packing correction. **(A)** T-stacking. **(B)** Parallel stacking. r_1: distance between the centers in aromatic planes of two side chains; r_2: horizontal displacement between the aromatic planes; r_3: distance from the center in one aromatic plane to the other aromatic plane; θ: dihedral angle between the aromatic planes (0 ~90 °).

In Eq. 2.8, the π-π packing correction is expressed as a function of distances r_1, r_2, and r_3 as well as dihedral angle θ (Figure 2.3).

$$E_{packing} = \sum C \cdot f(r_1) \cdot f(r_2) \cdot f(r_3) \cdot f(\theta) \qquad \text{(Eq. 2.8)}$$

$$f(x) = \exp\left[-A \cdot \exp\left[-B \cdot (x - x_0) \cdot (x - x_1)\right]\right] \qquad \text{(Eq. 2.9)}$$

f(x) is a normalized continuous function which shape is similar to a step function. In Eq. 2.9, coefficients A and B are constants with values 2.0, while x_0 and x_1 represent the boundaries for the variable x which values are shown in Table 2.2. The coefficient C for the packing correction is 3.0 kcal/mol, which is in the range of Jurecka's estimation.

Table 2.2. Geometry parameters for π-π packing correction.

	r_1 (Å)	r_2 (Å)	r_3 (Å)	θ (°)
x_0	3.5	0.0	0.0	0.0 (θ < 45°), 80.0 (θ >= 45°)
x_1	6.5	3.5	5.0	20.0 (θ < 45°), 90.0 (θ >= 45°)

Self-Contact Correction

It is common in proteins to find side chains of Asn, Gln, Ser, and Thr interacting with their own backbone nitrogen or oxygen atoms. Such an interaction depends on both the side chain conformation and the secondary structure, so that it is more complicated than a normal hydrogen bond. It was found by Pal *et al.* that self-contact interactions have specific roles in protein local environments,[98] while most of them have tertiary interactions, saturated by hydrogen bonds from side chain-backbone and side chain-solvent. Therefore, self-contact interactions are considered as a special case of hydrogen bonding in the VSGB 2.0 model.

The correction is represented as a sum of Gaussian functions dependent on the distance r between two heavy atoms with self-contact interactions (Eq. 2.10).

$$E_{self-contact} = \sum A \cdot \exp[-B \cdot (r - r_0)^2] \cdot C \qquad \text{(Eq. 2.10)}$$

One atom is the polar atom from the side chain of Asn, Gln, Ser, or Thr, while the other one is the backbone nitrogen or oxygen atom in the same residue. Based on our statistical study with high resolution PDB structures, Asn, Ser, and Thr are most likely to form self-contact interactions, so that the correction is stronger for these three amino acid residues. The value of r_0 was taken from the most populated distance for the corresponding amino acid residue. Ser and Thr are treated with identical parameters due to their similarity in side chain hydroxyl group. It is worth mentioning here that the coefficient C is used to rescale the correction if the amino acid residue is in a regular secondary structure, since the self-contact interaction disturbs the hydrogen bonding network in a helix or sheet. Parameters for the self-contact correction are presented in Table 2.3.

Table 2.3. Parameters of self-contact correction.

	Asn	Gln	Ser, Thr
A (kcal/mol)	-4.7	-2.5	-4.8
B	3.2	4.0	3.2
C	1.0	0.0	0.8
r_0 (Å)	3.2	3.8	3.2

Hydrophobic Term

The hydrophobic term was introduced by us in a previous version of our energy model. The original term was taken from a scoring function employed in docking calculations, ChemScore.[66,99] The hydrophobic term rewards contacts between nonpolar heavy atoms and stabilizes hydrophobic contacts. As the effect of hydrophobic term is much smaller in one side chain than in a loop, in the VSGB 2.0 model we replaced the linear function with a polynomial and refit the parameters based on the loop predictions at lengths of 11-13 residues.

$$E_{hydrophobic} = coeff \cdot \sum_{ij} E^{ij}_{hydrophobic} \qquad \text{(Eq. 2.11)}$$

$$E^{ij}_{hydrophobic} = \begin{cases} 0.0 & (1 \le scale) \\ 0.25 \cdot scale^3 - 0.75 \cdot scale + 0.5 & (-1.0 < scale < 1.0) \\ 1.0 & (scale \le -1.0) \end{cases} \qquad \text{(Eq. 2.12)}$$

$$scale = 2.0 \cdot (r_{ij} - r_i^{vdw} - r_j^{vdw} - 2.0) / 3.0 \qquad \text{(Eq. 2.13)}$$

The coefficient in Eq. 2.11 was fit by a line search algorithm and the optimal value we obtained is 0.30. (Comparison of the linear function and polynomial is shown in Figure 2.10 in appendix.)

The hydrophobic term is intended to model the interaction of hydrophobic surfaces presented by various protein and ligand groups with water; when these groups make contacts (in the extreme case forming the hydrophobic core of the protein), unfavorable interactions with water are eliminated, and this effect drives hydrophobic packing. The atom-atom contact term described above represents an alternative to models which attempt to directly compute cavitation

and Van der Waals interactions of the solute atoms with the solvent. Our empirical investigations, initially described in ref. [66] but continued over the past several years with extensive experiments on large data sets, suggest that the model of Eq. 2.12 and 2.13 provides superior predictions as compared to more standard approaches penalizing exposed hydrophobic surface area, which are generally derived from small molecules in bulk solution, a very different situation from hydrophobic groups embedded in an active site cavity, or otherwise positioned in confined spaces within a protein environment. Therefore, in the VSGB 2.0 model, we have eliminated the G_{cav} and G_{vdw} terms discussed above in Eq. 2.2 and 2.3, and replaced them with the hydrophobic term presented in this section. The quality of results for long loop prediction, which involve no further adjustment of the energy model, will serve as an unbiased test of the validity of this approach.

Optimization of the VSGB 2.0 Model

In the development of the VSGB 2.0 model, we fit the methods and parameters to 2239 single side chain and 100 11-13 residue loop predictions. The goal of optimizing the parameters in the VSGB 2.0 model was not only to improve the results of single side chain or loop prediction, but also to give a more accurate physical description that is transferable to tackle practical problems such as homology model refinement. As mentioned in the introduction, fitting to large high-quality experimental data sets helps to capture the correct physics in proteins and reduces the risks of overfitting.

The VSGB 2.0 model was optimized by carrying out single side chain prediction. As mentioned previously, the prediction of a single protein side chain can be determined by the lowest energy conformation either without minimization (single point) or with minimization. We tried to maximize the performance of the VSGB 2.0 model on both single point and

minimization selection during the procedure to optimize our model. Thus the parameter optimization is based on the ability to select the lowest energy structure, as well as the ability to make a good approximation of the energy funnel. Except the hydrophobic term which was fit to 11-13 residue loop predictions with decoys, all the parameters of the solvent model and physics-based corrections were optimized by a script based on a Monte-Carlo (MC) algorithm. The MC script generates parameters, recalculates the single point energy for each side chain candidate in the pool, and determines the prediction by the lowest energy candidate. We selected two sets of optimal parameters with the lowest average RMSDs suggested by the MC script, and performed the actual side chain predictions to determine the final parameters as the ones leading to the highest success rate with and without minimization.

2.3. Results

Single Side Chain and Loop (11-13 residues) Prediction

The results of single side chain prediction obtained by the VSGB 2.0 model are shown in Table 2.4 in comparison to our previous energy model (herein referred to as VSGB 1.0), which employs the OPLS-AA energy function with the original variable dielectric solvent model and the earlier implementation of the hydrophobic term. 11 polar or charged amino acids were included for the fitting in order to compare to the results from previous work. The overall accuracy of the 2239 single side chain predictions is as high as 93.0% with single point energy evaluation (SP) and 91.6% with minimized energy evaluation (MIN) respectively. Generally the VSGB 2.0 model improves the prediction accuracy for 1.0% (SP) and 2.0% (MIN). The most significant improvements come from Gln and Asn, at 3.1% (SP), 5.6% (MIN) and 2.0% (SP), 3.2% (MIN). Another remarkable improvement advanced by the VSGB 2.0 model is to reduce the large differences in accuracy between single point energy evaluation and minimized energy

evaluation. With the VSGB 1.0 model, the difference in success rate is 7.5% for Lys and 8.1% for Gln; with the VSGB 2.0 model, it is reduced to 1.6% and 5.6% respectively. Given the high success rate of the VSGB 1.0 model, such an improvement is non-trivial. It is evident that the VSGB 2.0 model improves not only the energy score but also the gradient of the energy score.

Table 2.4. Summary of single side chain prediction results for 11 polar or charged residues.

		VSGB 2.0		VSGB 1.0		Zhu *et al.*[b]	
Residue Type	Number of Cases	SP (%)[a]	MIN (%)[a]	SP (%)[a]	MIN (%)[a]	Number of Cases	SP (%)[a]
Arg	144	85.4	84.0	83.3	82.6	171	77.8
Asn	252	92.5	91.7	90.5	88.5	237	85.7
Asp	293	94.9	94.9	94.2	92.5	254	91.7
Cys	92	100.0	100.0	100.0	100.0	49	93.9
Gln	161	88.8	83.2	85.7	77.6	161	85.7
Glu	152	88.8	86.2	88.2	84.9	193	79.3
His	83	95.2	95.2	91.6	91.6	132	86.4
Lys	121	91.7	90.1	95.9	88.4	198	76.8
Thr	404	95.8	94.3	94.8	92.6	302	92.4
Tyr	221	99.5	99.1	99.1	98.6	184	89.7
Ser	316	88.9	88.0	88.3	86.1	297	79.1
All	2239	93.0	91.6	92.0	89.6	2178	85.0

All the single side chain predictions were performed with ICDA prepared structures.
[a] A "successful" prediction is defined as one where the heavy atom RMSD < 1.5 Å to the native side chain conformation. The percentages reported in here are the ratio of the number of accurate predictions to the total number of predictions. "SP" stands for the single point energy evaluation; "MIN" stands for the minimized energy evaluation.
[b] This method uses the original single side chain prediction algorithm, a different data set and the VSGB 1.0 model. (See ref. [68])

The hydrophobic term was optimized with loop predictions of 11-13 residue length, results of which are presented in Table 2.5. Although the change from linear form to polynomial form did not significantly improve the accuracy, we still believe that the basic physical arguments can easily justify our reformulation, which here allows for the introduction of a continuous first derivative and the uniform application of the term throughout all of space. In

particular, a continuous first derivative would take the force into consideration and allow a smoother minimization. Likewise, to our knowledge, there is no experimental evidence to suggest that the hydrophobic forces exerted between crystal mates should be any different than those hydrophobic forces exerted within the unit cell. Thus, the parsimony principle seems to motivate the change, even in the absence of strong data suggesting the polynomial form adopted here is fundamentally more accurate than the earlier linear ramping function.

Table 2.5. Summary of 11-13 residue loop prediction results with different forms of hydrophobic term.

Hydrophobic term	Included the crystal environment	Average Backbone RMSD (Å)	Accuracy (%)
Linear	No	0.81	91.0
Polynomial	No	0.83	91.0
Polynomial	Yes	0.85	91.0

It should be noted that the average RMSD of ~0.8 Å reported in Table 2.5 arises from selection of loops from a suite of decoys generated with an older energy model from ref. [66], as opposed to full scale optimization of loop prediction with the current energy model and sampling algorithms. In view of the results to be reported below for longer loops, it is likely that improved results would be obtained for 11-13 residues loops with the latter, more rigorous sampling protocol.

Super Long Loop (14-20 residues) Prediction

In order to validate the effectiveness of the VSGB 2.0 model, a super long loop set was used to test the accuracy of the energy function described herein. A summary of the results is shown in Table 2.6 and Figure 2.4, while the detailed results are given in Table 2.13 in appendix.

Table 2.6. Summary of super long loop prediction results.

Loop Length	Number of Cases	VSGB 2.0 with ICDA				VSGB 1.0 without ICDA			
		Median Backbone RMSD (Å)	Average Backbone RMSD (Å)	Average Side Chain RMSD (Å)	% of Cases with RMSD < 2Å	Median Backbone RMSD (Å)	Average Backbone RMSD (Å)	Average Side Chain RMSD (Å)	% of Cases with RMSD < 2Å
14	36	0.38	0.51	1.67	100.0	0.67	1.19	2.51	91.7
15	30	0.54	0.63	1.85	100.0	0.75	1.55	3.07	73.3
16	14	0.43	0.70	1.85	100.0	0.80	1.43	3.20	78.6
17	9	0.57	0.62	1.84	100.0	1.92	2.30	4.25	66.7
18	16	0.60	0.80	1.78	100.0	3.45	4.18	5.59	37.5
19	7	1.60	1.41	3.46	100.0	1.31	2.65	3.87	57.1
20	3	1.68	1.59	2.88	100.0	1.12	1.43	2.71	66.7
All	115	0.52	0.69	1.91	100.0	1.04	1.89	3.37	73.0

RMSD calculation: A predicted structure is superimposed to the native protein structure excluding the target loop. Backbone RMSDs are calculated with N, C_a and C atoms; side chain RMSDs are calculated with heavy atoms.

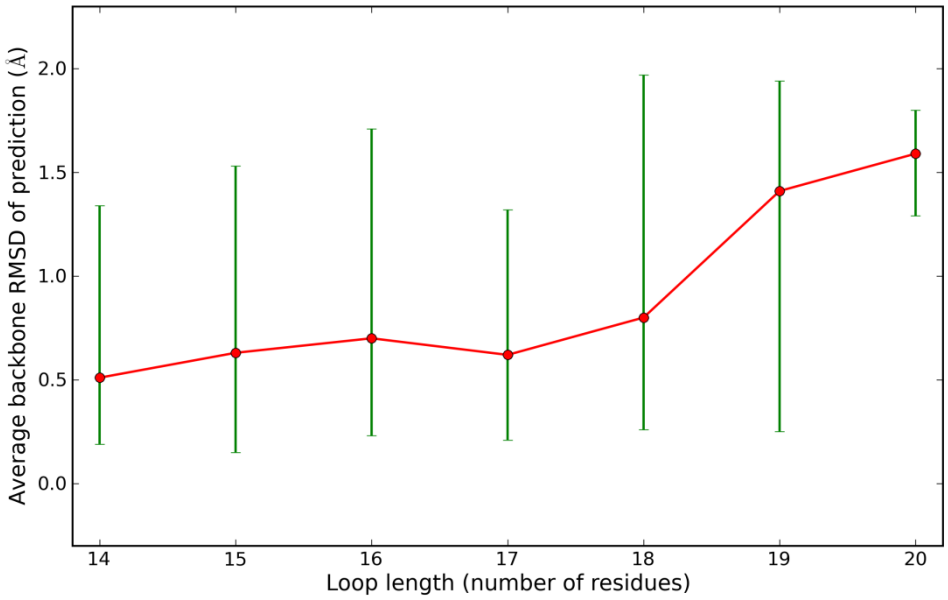


Figure 2.4. Accuracy of super long loop predictions. Red dots represent the average backbone RMSDs while green lines represent the ranges of the backbone RMSD.

Testing on the same data set, we are able to make a direct comparison of the relative performance of the VSGB 2.0 model and the VSGB 1.0 model. Using the VSGB 2.0 model, 100.0% of loop predictions have backbone RMSDs below 2.0 Å from the native conformations,

where only 73.0% of the predictions made with the VSGB 1.0 model would be similarly accurate. Furthermore, the better performance of the VSGB 2.0 model was independent of any particular chosen success criteria, as depicted by Figure 2.5, where the percentage of predicted loops with RMSD lower than a cutoff is plotted as a function of the given cutoff. The prediction difficulty often increases dramatically with loop length; however, our results with the VSGB 2.0 model reflect average backbone RMSDs in a narrow range from 0.51 to 0.80 Å for 14-18 residue loops. Even though it is extremely challenging to predict loops longer than 18 residues, the VSGB 2.0 model still displays high accuracy with average backbone RMSDs of only 1.41 Å and 1.59 Å for 19 and 20 residue loops. Apart from the backbone, the VSGB 2.0 model also significantly reduces the RMSDs for side chains in the loops. For example, for 18 residue loops, we obtained average backbone/side chain RMSDs of prediction are 0.80/1.78Å with the VSGB 2.0 model compared to 4.18/5.59Å with the VSGB 1.0 model.

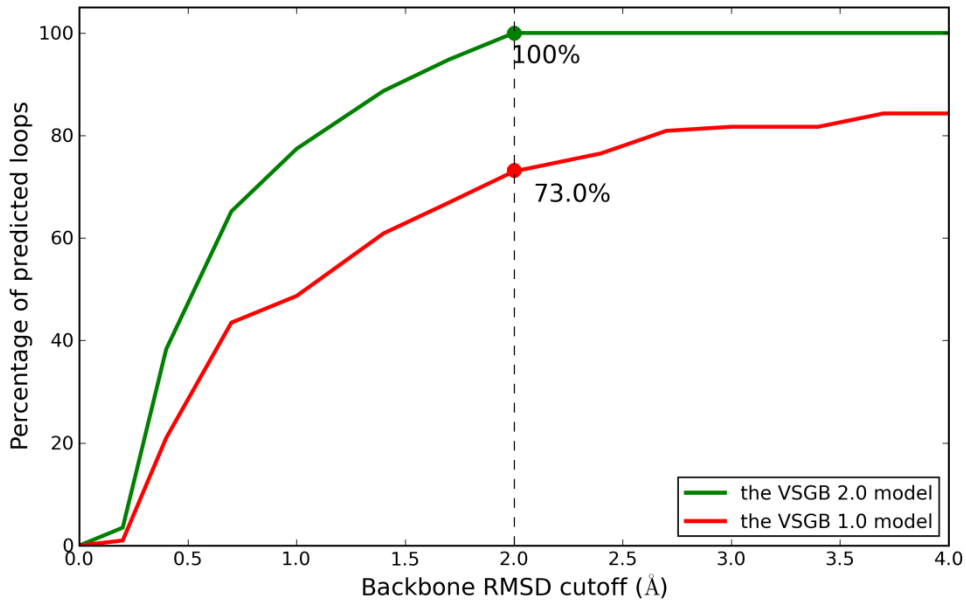


Figure 2.5. Comparison of loop prediction results using the VSGB 1.0 and 2.0 models. The y-axis shows the percentage of predicted loops within the RMSD cutoff on the x-axis.

For super long loop prediction, the hierarchical algorithm is able to enrich the native-like conformations along the increasing fixed stages. The analysis of the average RMSDs at each stage shows that the VSGB 2.0 model not only provides an accurate score for the final prediction, but also improves the accuracy for each prediction stage (Figure 2.6), allowing higher percentage of native-like confirmations and faster convergence towards the global minimum. For example, both 14 and 15 residue loop predictions reach average RMSD below 2.0 Å at least one stage earlier with the VSGB 2.0 model. This also implies that fewer sampling stages are required with the VSGB 2.0 model, making the loop predictions more cost efficient.

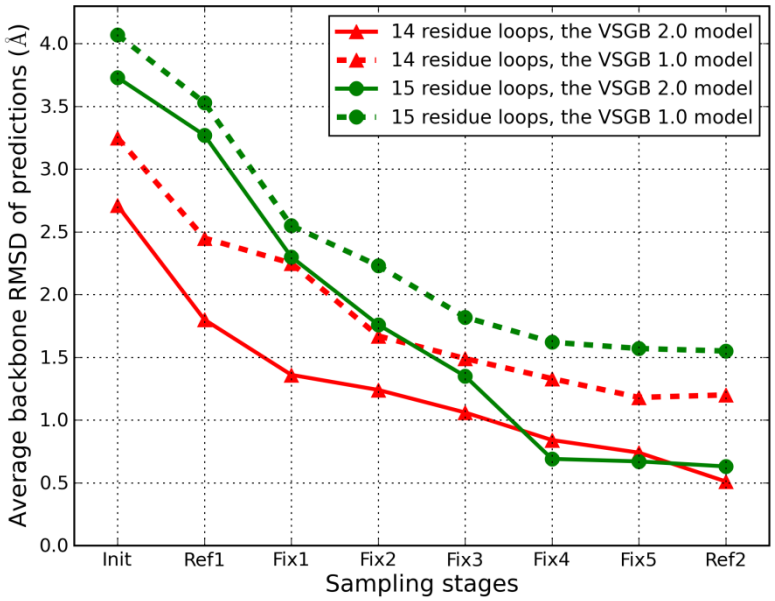


Figure 2.6. Average backbone RMSD of the predicted 14 and 15 residue loops at each stage.

Super Long Loop (14-20 residues) Predictions in Inexact Environments

A subset of 10 cases ranging from 14 to 16 residues (all with reliable well resolved electron densities for the target loop as well as surrounding side chains) were used to test loop predictions in inexact protein environments. An inexact environment was here created by replacing the target loop by a non-native loop conformation (RMSD>3.0 Å) and minimizing the

new structure with surrounding side chains. The loop prediction in inexact environments was performed with method SLLP-SS, as was described earlier in Methods and Materials section. The test results are shown in Table 2.7.

Table 2.7. Results of super long loop prediction in inexact environments.

				exact environment			**inexact environment**		
PDBID	Loop start	Loop length	Number of residues surroundi-ng	Starting structure RMSD (Å)	Backbone RMSD (Å)	Side chain RMSD (Å)	Starting structure RMSD (Å)	Backbone RMSD (Å)	Side chain RMSD (Å)
1E6U	A274	14	28	0.00	0.27	0.85	3.37	0.31	1.07
1ZEQ	X53	14	61	0.00	0.21	2.10	3.52	0.28	2.08
2BWR	A269	14	69	0.00	0.31	1.70	3.35	0.58	2.64
3BY9	A205	14	59	0.00	0.32	0.81	3.14	0.41	0.83
3EHR	A95	14	83	0.00	0.51	3.40	4.09	1.30	4.08
1QAZ	A298	15	51	0.00	0.99	3.00	3.46	2.18	4.52
1RA0	A283	15	36	0.00	0.30	2.13	6.51	1.24	2.38
1RA0	A361	15	81	0.00	0.52	1.66	5.10	0.52	1.70
3CSS	A95	15	52	0.00	0.52	1.39	3.06	1.98	3.62
1WM3	A67	16	49	0.00	0.23	1.24	3.08	0.26	1.63
Average RMSD					0.42	1.83		0.91	2.47

The RMSD of a starting structure was calculated as the backbone RMSD of the target loop in the starting structure, but the prediction of the loop was actually built from scratch.

Compared to the starting structures, all the 10 cases display improvements in RMSDs after the loop reconstruction. This confirms that the VSGB 2.0 model together with the augmented sampling method presented herein is able to improve the quality of models, suggesting a high potential to succeed in refining comparative models. Although starting from the inexact environment degrades the overall performance by 0.49 Å in the backbone RMSD, many of the cases actually reach similar accuracy as starting from the exact environment, such as 1ZEQ X53-66 and 3BY9 A205-218. Furthermore, the average error is still a highly satisfactory 0.91 Å, and the maximum RMSD remains less than 2.5 Å.

2.4. Discussions

The single side chain prediction results represent a major advance as compared to the accuracy levels reported in ref. [68]. A great deal of the improvement is due to the use of a better side chain data set (in which all of the atoms in the side chain are reliably located by from the electron density obtained from the crystallographic observations), as opposed to improvements in the energy model. These results demonstrate that data set quality is vital in both development and assessment of molecular models. Problems with the data set will invariably produce misleading (and in some cases highly misleading) conclusions with regard to the performance of the model.

The improvements in accuracy attributable to the new model are relatively small, but significant, particularly as they are clustered in the polar and charged side chains. Furthermore the vast majority of cases, where predictions failed, have quite small energy gaps, implying that the impact on loop prediction of such errors would be relatively minimal. However, we note that elimination of erroneous predictions is not the only benefit to adding a new, physically important term to the energy model. In many cases, the VSGB 1.0 model may have given a correct prediction for a particular side chain, but not properly evaluated the energy of the side chain conformation as compared to possible alternatives. As an example, consider the π-π packing term which rewards stacking of aromatic residues. In many cases, these residues are in the hydrophobic core of the protein, and re-prediction of the side chain conformation of the residue can have only one outcome due to steric considerations; the problem is like a jigsaw puzzle in which the “piece” can fit back into the puzzle in only one “conformation”. However, the energy gained from forming the entire core in the specific fashion that enables many stacking interactions to be formed may be underestimated by an energy model that does not reward stacking interactions. This would have consequences not only for loop prediction (where, as we show below, incorrect loop conformations often fail to make key stacking interactions present in

the native conformation) but also for refining a homology model so that it has the optimal interlocking pattern of side chain interactions. Optimization of the π-π stacking and other terms to improve side chain prediction uses a small fraction of "sensitive" side chain cases to detect problems in the energy model and optimize the terms that improve these cases. The success of this strategy is manifest in the results reported above for long loop prediction, where very large improvements in average backbone RMSD and associated average side chain RMSD are seen uniformly for the 14-20 residue loop database. The VSGB 2.0 model fixes a number of cases which previously had substantial energy errors leading to inaccurate loop RMSDs.

Computational Cost of Single Side Chain and Super Long Loop Predictions

One single side chain prediction (without clustering) usually takes from a few seconds (e.g. for His and Cys which have only a few rotamers) to 1 hour (e.g. for Lys and Arg which have thousands of rotamers) with a single 2.66G Hz Intel Xeon processor. With the same architecture, the average time needed in super long loop prediction for the stage of initialization, the stage of refinement, and one stage of fix sampling is 10, 30, and 100 hours respectively. Since all the calculations in one stage can be carried out in parallel, a full loop prediction with sampling up to "Fix 5" takes around 1 day to finish with 20 processors.

Importance of Systematic Application of Protonation State Assignment

When making either single side chain predictions or loop predictions, it is extremely difficult to achieve accurate predictions if one or more ionizable side chains are represented in the incorrect protonation state. As a simple example, some Asp and Glu residues form carboxylate "dimers" with neighboring Asp or Glu residues without any nearby metal ions; in the crystal structure, oxygen atoms from the pair of carboxylates belonging to each side chain are

observed to be within hydrogen bonding distance (~3.0 Å). This type of structure implies that at least one of the carboxylates must be protonated. This costs free energy with regard to the standard protonation state of a carboxylate in solution, but is clearly necessary to avoid large repulsive interactions between charged oxygen atoms that would otherwise occur (vs. forming a strong hydrogen bond). If the unprotonated forms are used, the "dimer" structure will never be predicted as lowest in energy. Loop prediction is more subtle, but side chains in loops do form salt bridges and hydrogen bonds which are dependent upon protonation state. If these interactions cannot be formed due to failure to incorporate a nonstandard (but accessible) protonation state, the energy of the native loop conformation may fail to be competitive with incorrect alternatives.

All of the loop and side chain prediction results shown below have been generated by running an automated protonation state assignment program, the ICDA, based on methods described in ref. [67], on the entire protein of each member of the test set. Since the publication of ref. [67], we have put significant efforts into improving the ICDA by running a large number of test cases and examining them visually to make sure that obvious errors are eliminated. However, the data sets in the present paper have been treated in an automated fashion. The effects on accuracy of failing to run the ICDA has been examined for a selected set of loops, those which yielded large energy errors in previous work which did not employ systematic ICDA preparation (but also used a different energy model). By rerunning these cases with ICDA preparation and the VSGB 1.0 model, it is possible to identify a subset of cases where ICDA preparation is essential to achieving accurate results. At least 8 of 115 cases (7.0 %) require ICDA preparation to achieve accurate structure prediction. (Table 2.8)

Table 2.8. Cases that require ICDA preparation to achieve accurate prediction.

PDBID	Loop Start	Loop Length	VSGB 2.0, with ICDA		VSGB 1.0, with ICDA		VSGB 1.0, without ICDA	
			EGAP (kcal/mol)	Backbone RMSD (Å)	EGAP (kcal/mol)	Backbone RMSD (Å)	EGAP (kcal/mol)	Backbone RMSD (Å)
1RA0	A283	15	7	0.30	0	1.03	-39	2.78
2PKF	A26	15	0	0.65	10	0.47	-8	2.34
2BG1	A708	16	-3	0.40	5	0.59	-4	2.15
2PYW	A321	16	-1	0.99	-5	0.72	-37	2.57
2HDW	A131	17	10	0.60	-3	0.42	-126	2.22
3CUZ	A384	18	4	0.82	-4	0.70	-22	3.45
3EH1	A813	18	11	1.60	5	1.54	-22	2.84
3GGQ	A550	18	8	0.37	8	0.53	-26	5.12

EGAP: energy of the prediction – energy of the minimized native structure.

In the present work, we are able to assign protonation states with high accuracy due to having the native structure available. In the context of realistic homology model refinement, the native structure would not be known in advance. Achieving the correct protonation states would then require sampling differing protonation states on the fly during the simulations, and/or running a number of alternative protonation states as a part of the iterative refinement algorithm protocol.

One interesting test towards the development of on-the-fly protonation state sampling algorithm is to predict the *p*H-dependent conformational changes. One good example is the bovine β-lactoglobulin, which X-ray crystallography has distinguished its different conformations of loop 85-90 at *p*H 6.2 and 8.2.[100] A preliminary test showed that at different assigned *p*H values, ICDA correctly assigned the protonation states of Glu89 which is buried at *p*H 6.2 (protonated, PDBID: 3BLG) and exposed to solvent at *p*H 8.2 (deprotonated, PDBID: 2BLG). With the correct protonation state assignment, our loop prediction had a good agreement with the crystal structures (backbone RMSD 1.3 Å and 0.7 from native structures at *p*H 6.2 and

8.2). (Figure 2.7) This test has indicated the capability of ICDA in correctly predicting the protonation states with different conformations, and can be interpreted as a progress in the development of on-the-fly protonation state assignment.

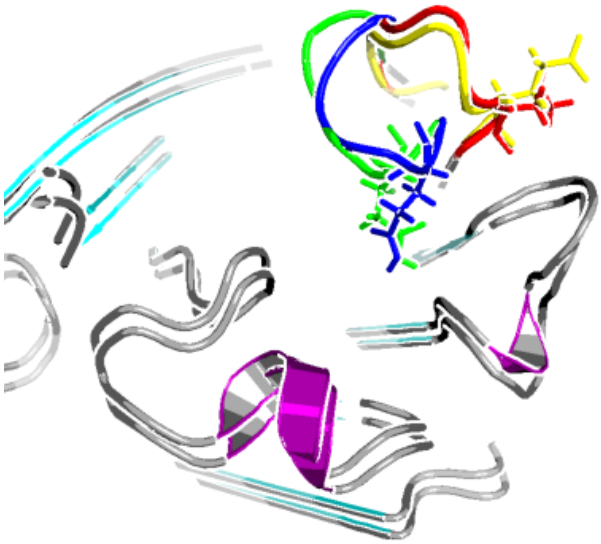


Figure 2.7. Overlay of loop predictions (loop A84-91) at *p*H 6.2 and 8.2 to their corresponding native structures for bovine β-lactoglobin. At *p*H=6.2, prediction (green, backbone RMSD=1.3 Å) and native (blue, PDBID: 3BLG); at *p*H=8.2, prediction (red, backbone RMSD=0.7 Å) and native (yellow, PDBID: 2BLG).

Further, a number of publications in the literature describe successful on-the-fly protonation state sampling algorithms, which could be adapted to our refinement algorithms. It is likely though that a significant effort will have to be put into making sure that, these methods are sufficiently accurate to reliably achieve refinement objectives. The results presented in this dissertation should be viewed as proof of concept, demonstrating that if accurate protonation states can be assigned, significant improvement in blind structural prediction efforts will result. Running tests without proper protonation states, in contrast, would make it impossible to distinguish intrinsic failures of the energy model from failures to assign the correct protonation state.

A Better Description of Protein Energy Landscape

As a mini folding problem, super long loop prediction at high resolution demands an energy model with highly precise physical description of the loop in question as well as of the environment. However, an inaccurate energy model with missing physics usually fails to discriminate the native conformations from the non-native ones, leading to wasted sampling effort in the false global minimum which could be far away from the true one. This explains why sometimes a more intensive sampling effort leads to a poorer prediction. Through our analysis of such mispredictions, the missing physics in our previous energy models appeared to be stemming from inaccurate descriptions of electrostatic, hydrogen-bonding, hydrophobic, and π-π interactions..

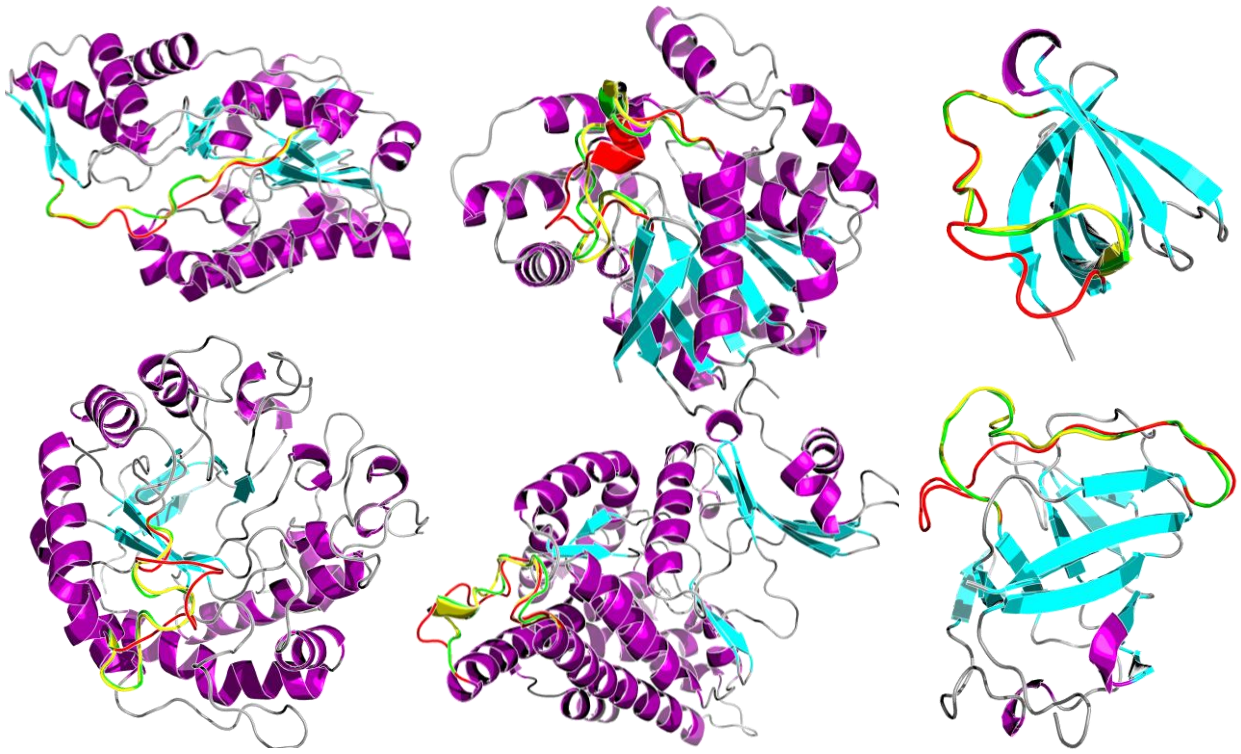


Figure 2.8. Overlay of 6 predicted loops using VSGB 1.0 (red) and VSGB 2.0 (green) models to their native structures (yellow).

Using the optimized variable dielectric solvent treatment and the additional physics-based corrections, the VSGB 2.0 model is likely to provide a better, more complete physical description for protein high resolution modeling. The addition of corrections, especially to the hydrogen bonding, π-π interactions and hydrophobic interactions, compensates the effects that were incompletely described by the electrostatics, Van der Waals, and nonpolar interactions in the VSGB 1.0 model. As a result, the corresponding native or native-like conformations are stabilized, with respect to the competing non-native conformations, and thus are more likely to be the global minimum energy structure on the potential energy surface. In addition to the more accurate description of the global minimum energy structure, the VSGB 2.0 model also incorporates a stronger bias along the whole of the energy surface towards more native-like conformations, which could be due to the improved physical description of the overall energy surface. One indication of the greater bias of potential energy surface of the new energy function towards more native like structures is shown in Figure 2.8, where the VSGB 2.0 model gives lower average backbone RMSDs for each stage compared to the VSGB 1.0 model. This indication could be interpreted as preliminary evidence that the new energy function is better describing the "folding funnel" of the protein potential energy surface.[101,102] Another indication, more direct, is the consistent correlations of relative energies and RMSDs considering all the conformations that have been sampled during the loop prediction. An example is presented in Figure 2.9 to further demonstrate the better physical description in the VSGB 2.0 model using the RMSD-energy correlation.

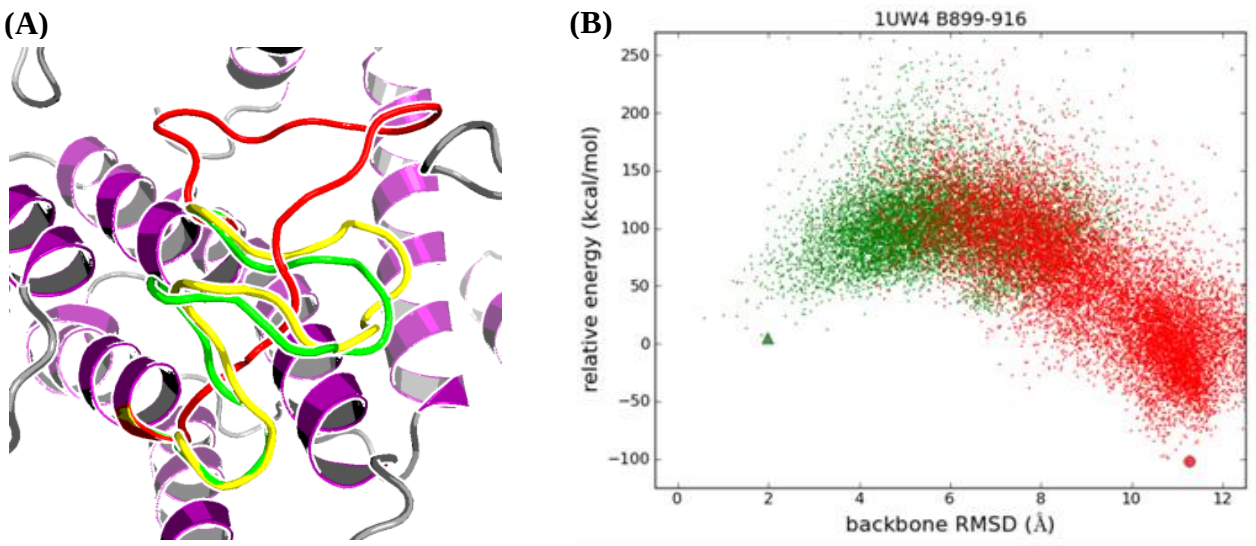


Figure 2.9. Loop prediction results with serious energy error in the VSGB 1.0 model fixed by VSGB 2.0 model. **(A)**: overlay of the native (yellow) structure and the prediction using VSGB 1.0 (red) and VSGB 2.0 (green) models.PDBID: 1UW4. Loop B899-916: backbone RMSD =2.0 Å, EGAP=5 kcal/mol (VSGB 2.0); backbone RMSD=11.3 Å, EGAP=102 kcal/mol (VSGB 1.0). **(B)**: relative energies (VSGB 2.0: green; VSGB 1.0: red) versus backbone RMSDs.

Water between Protein Molecules in Crystal Structure

The 14 residue loop (A153-166) in a nucleotidase (PDBID: 1JP4) is unique in our data set of super long loops: at the beginning, neither the VSGB 1.0 nor the VSGB 2.0 model alone gave reasonable predictions. (The VSGB 1.0 model: backbone RMSD=7.26 Å, EGAP=43 kcal/mol; the VSGB 2.0 model: backbone RMSD=2.40 Å, EGAP=30 kcal/mol). Considering our starting point of improving the physical description, what is the missing physics in this case? An analysis of the native structure has shown that the water molecules bridge the loop in question and the crystal environment, which could lead to the incorrect energy evaluation due to the limitations of the implicit solvent model.

The sequence of this 14 residue loop (PYYNYQAGPDAVLG) has a high fraction of non charged residues, which in this case have strong hydrophobic interactions, unusually, predominantly with several other protein molecules in the crystal environment, as opposed to

with the hydrophobic core of the protein molecule which the loop is a part of. Instead of forming the extended native conformation which contacts the neighboring proteins, all the mispredictions are packed into the protein body. However, we found that there are three bound water molecules (HOH 858, 1057, and 1083) which contribute significantly to the stability of the extended native conformation: all these water molecules connect the loop to the crystal environment through their hydrogen bonding network and consequently pin down the extended conformation (Figure 2.10). Such a first-shell solvation effect is unlikely to be well represented by any implicit solvent model, and thus we could not create the correct physical environment unless these water molecules were included.

In order to repredict this loop in the correct physical environment, we explicitly added these three water molecules and the ones that form hydrogen bonds with them (HOH 858, 911, 1057, 1083, 1108, 1145, and 1173) to our all-atom model. Hydrogen bonds between the explicit water molecules and the proteins were considered. With the presence of these water molecules and the hydrogen bonding correction applied to protein-water interactions, the prediction with the VSGB 2.0 model yields high accuracy (backbone RMSD=0.31 Å, side chain RMSD=0.67 Å, EGAP=4 kcal/mol). This case study highlights the limitations of implicit solvent models, and suggests possible treatments of including or predicting crystal water positions in future works.

It should be noted that the explicit waters are required specifically in the interstitial region between protein molecules in the crystal structure. For a single protein molecule in solution, the loop in question would in fact almost certainly adopt a different conformation, since the hydrophobic region of the loop is buried in a hydrophobic region of the neighboring protein molecule in the crystal; in fact, it might well be found in the competing conformation selected

without the crystal waters, in which the hydrophobic interactions of the loop are primarily intramolecular, as opposed to with the neighboring molecule in the crystal.

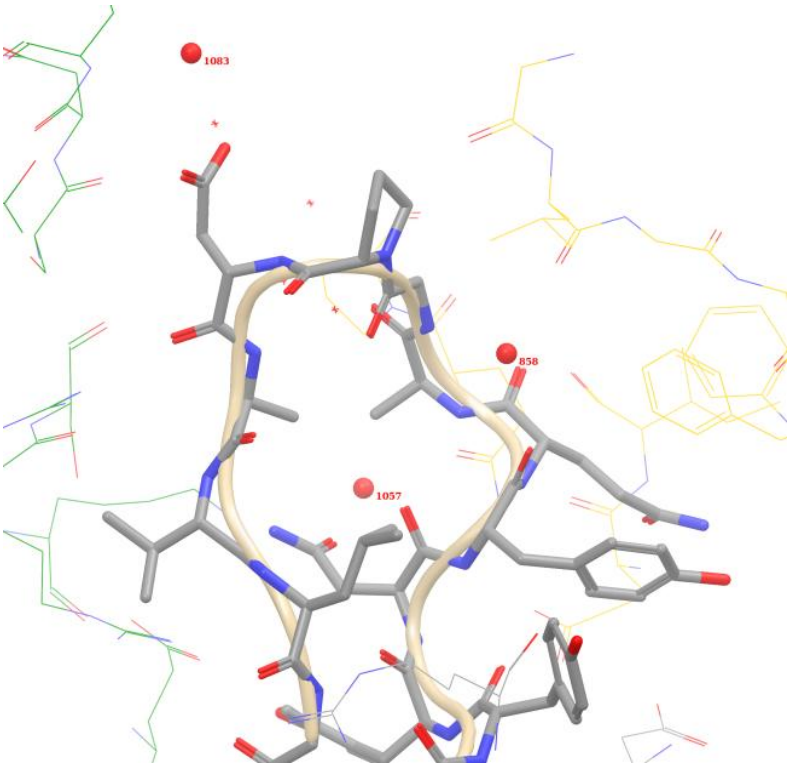


Figure 2.10. The 14 residue loop in a nucleotidase (PBDID: 1JP4. Loop A153-166) and bound crystal water molecules. Protein molecules from the crystal environment are shown in line representation with green or yellow carbons. Bridging water molecules are shown in red spheres.

Possible Energy Errors in Current Data Set

The simulation of 115 long loops carried out in this paper requires a substantial amount of computation time. Furthermore, the lengthiest simulations (using 10 fixed stages, referred to in the text as "Fix 10") requires considerably more computational effort than the 5 fixed stage ("Fix 5") algorithm that was employed for most of the calculations. This is the reason that in the data set presented above we utilized the "Fix 10" algorithm only in cases where the "Fix 5" simulations displayed an RMSD greater than 2.0 Å.

However, it must be recognized that such a protocol could conceivably lead to bias in the results, as we do not know whether running the full "Fix 10" protocol on all of the cases would produce additional errors. We have carried out an assessment of this problem, and cases identified with energy errors using "Fix 10" sampling are listed in Table 2.9.

Table 2.9. Cases with energy error from "Fix 10" sampling.

PDBID	Loop start	Loop length	"Fix 5"		"Fix 10"	
			EGAP (kcal/mol)	Backbone RMSD (Å)	EGAP (kcal/mol)	Backbone RMSD (Å)
1N0Q	A24	14	-12	0.51	-18	3.62
1RA0	A361	15	-5	0.52	-9	3.47
3CSS	A95	15	-17	0.52	-24	2.77
2PYW	A321	16	-1	0.99	-10	4.45
3CUZ	A384	18	4	0.82	-11	2.77
1NC5	A193	19	-12	1.79	-23	3.61

2.5. Conclusions

In this work, we have described a new energy model (VSGB 2.0) which contains an optimized solvent model and physics-based correction terms. The VSGB 2.0 model was fit to a large database of protein single side chain and loop (11-13 residues) prediction and validated by a large set of super long loop predictions. It was shown that the VSGB 2.0 model, combined with the systematic protonation state assignment, improved the accuracy of super long loop predictions by 27.0% compared to our previous energy model (VSGB 1.0). A series of extensive analysis shows that the VSGB 2.0 model not only improved the results of single side chain and loop predictions, but also provides a better physical description for high resolution protein structure modeling. Further tests will include receptor-ligand docking, longer loop prediction, loop-helix-loop prediction, and applications on a variety of proteins, such as kinases, G protein-coupled receptors (GPCRs), and cytochrome P450s (CYP).

2.6. Appendix for Chapter 2

Table 2.10. Proteins in the single side chain set.

No.	PDB ID	Resolution (Å)	pH	Length (residues)
1	1A6M	1.00	7.0	151
2	1BYI	0.97	6.5	224
3	1C7K	1.00	7.0	132
4	1EB6	1.00	7.5	177
5	1G6X	0.86	7.5	58
6	1GQV	0.98	6.5	135
7	1IQZ	0.92	6.5	81
8	1L9L	0.92	7.0	74
9	1LNI	1.00	7.2	96
10	1MN8	1.00	7.5	100
11	1MUW	0.86	7.5	386
12	1N1P	0.95	7.4	504
13	1NKI	0.95	7.0	135
14	1NLS	0.94	6.8	237
15	1NWZ	0.82	7.0	125
16	1OAI	1.00	6.5	68
17	1PJX	0.85	6.5	314
18	1R6J	0.73	7.0	82
19	1SSX	0.83	8.0	198
20	1UFY	0.96	7.0	122
21	1UG6	0.99	7.4	431
22	1VYR	0.90	6.2	364
23	1W0N	0.80	6.5	131
24	1X8P	0.85	7.4	184
25	1ZLB	0.97	8.0	122
26	1ZZK	0.95	8.0	82
27	2A6Z	1.00	6.3	222
28	2B97	0.75	7.3	71
29	2BF6	0.97	7.0	449
30	2FOU	0.99	7.7	260
31	2FVY	0.92	7.8	309
32	2GGC	1.00	7.0	263
33	2H3L	1.00	7.5	103
34	2IIM	1.00	7.5	62
35	2IXT	0.80	7.5	310
36	2P5K	1.00	6.5	64
37	2PND	1.00	7.5	119

38	2PPN	0.92	7.0	107
39	2PVB	0.91	8.0	108
40	2QSK	1.00	8.0	95
41	2QXI	1.00	6.5	224
42	2VHK	0.94	7.1	206
43	3EA6	0.92	7.0	219
44	3FSA	0.98	7.0	125
45	3GOE	0.97	7.7	82

Table 2.11. Proteins in the loop set (11-13 residues).

No.	PDB ID	Resolution (Å)	pH	Length (residues)
1	1A8D	1.57	7.0	452
2	1AKO	1.70	7.0	268
3	1AOL	2.00	6.5	228
4	1AOZ	1.90	5.4	552
5	1ARB	1.20	7.4	268
6	1BHE	1.90	6.5	376
7	1BKP	1.70	6.5	278
8	1BN8	1.80	6.5	420
9	1BOL	2.00	6.7	222
10	1BX4	1.50	7.5	345
11	1C5E	1.10	6.5	95
12	1CB0	1.70	7.4	283
13	1CB8	1.90	6.5	678
14	1CNV	1.65	8.0	299
15	1CS6	1.80	7.5	382
16	1D0C	1.65	6.5	444
17	1DPG	2.00	5.6	485
18	1DQZ	1.50	7.8	280
19	1DYS	1.60	7.5	348
20	1ED8	1.75	7.5	449
21	1EDT	1.90	7.0	271
22	1EL5	1.80	7.0	389
23	1EOK	1.80	7.0	290
24	1EUR	1.82	7.0	365
25	1EXM	1.70	7.0	406
26	1F46	1.50	7.5	140
27	1FGK	2.00	6.5	310
28	1G6S	1.50	7.0	427
29	1G8F	1.95	7.5	511

30	1G9G	1.90	7.5	629
31	1GMU	1.50	6.5	143
32	1GPI	1.32	7.0	431
33	1GQV	0.98	6.5	135
34	1H4A	1.15	7.0	173
35	1HXH	1.22	7.5	253
36	1I4J	1.80	7.0	110
37	1I7P	2.00	7.5	274
38	1IIR	1.80	6.5	415
39	1IOO	1.55	6.5	196
40	1IU8	1.60	7.4	206
41	1IYE	1.82	7.5	309
42	1JP4	1.69	6.5	308
43	1KBL	1.94	7.0	873
44	1KCM	2.00	6.5	270
45	1KRH	1.50	7.0	338
46	1L8A	1.85	7.1	886
47	1LKI	2.00	7.0	180
48	1LMI	1.50	7.0	131
49	1M3S	1.95	6.5	186
50	1MLA	1.50	7.0	309
51	1MO9	1.65	6.5	523
52	1MS9	1.58	7.5	648
53	1MY7	1.49	7.5	114
54	1NLN	1.60	6.5	204
55	1NOG	1.55	7.5	177
56	1NSC	1.70	7.0	390
57	1O6L	1.60	7.5	347
58	1OCK	1.80	7.0	412
59	1OJQ	1.68	6.5	212
60	1OTH	1.85	7.5	321
61	1OYC	2.00	7.0	400
62	1P1M	1.50	7.0	406
63	1PGS	1.80	7.0	314
64	1PKH	1.42	7.0	204
65	1QLW	1.09	7.0	328
66	1QQP	1.90	7.5	328
67	1T1D	1.51	7.5	100
68	1WHI	1.50	7.0	122
69	1XYZ	1.40	7.0	347

70	2HLC	1.70	7.0	230
71	2PTD	2.00	7.5	298
72	2TGI	1.80	7.0	112

Table 2.12. Proteins in the super long loop set (14-20 residues).

No.	PDB ID	Resolution (Å)	pH	Length (residues)
1	1AH7	1.50	7.0	245
2	1BHE	1.90	6.5	376
3	1C1K	1.45	6.5	217
4	1DJ0	1.50	6.5	264
5	1E6U	1.45	6.5	321
6	1GPI	1.32	7.0	431
7	1GQ6	1.75	7.5	313
8	1H4A	1.15	7.0	173
9	1J83	1.70	7.0	180
10	1JP4	1.69	6.5	308
11	1JU3	1.58	7.5	585
12	1KWG	1.60	6.5	645
13	1N0Q	1.26	7.0	93
14	1NC5	1.60	7.5	373
15	1O97	1.60	6.5	584
16	1OCK	1.80	7.0	412
17	1ODM	1.15	7.5	331
18	1P3C	1.50	7.0	215
19	1P3D	1.70	7.5	475
20	1Q0R	1.45	7.5	298
21	1QAZ	1.78	7.5	351
22	1QL0	1.10	6.5	2.41
23	1QLW	1.09	7.0	328
24	1QQF	1.45	7.0	277
25	1R6D	1.35	6.5	337
26	1R6X	1.40	6.5	395
27	1RA0	1.12	7.5	430
28	1RDQ	1.20	8.0	370
29	1RQW	1.05	6.8	207
30	1RV9	1.53	7.0	259
31	1RYO	1.20	7.7	327
32	1S95	1.60	7.5	333
33	1UG6	0.99	7.4	431
34	1UW4	1.95	6.5	91

35	1V8H	1.20	6.1	107
36	1VJU	1.40	6.5	309
37	1VYR	0.90	6.2	364
38	1WB4	1.40	7.5	297
39	1WHI	1.50	7.0	122
40	1WM3	1.20	8.0	72
41	1WRJ	2.00	7.5	156
42	1WUI	1.04	7.4	156
43	1XFK	1.80	7.3	336
44	1XU1	1.90	7.5	138
45	1Y12	1.95	7.5	165
46	1YW5	1.60	7.5	177
47	1ZEQ	1.50	7.5	84
48	1ZHV	1.50	6.8	134
49	1ZHX	1.50	6.5	438
50	2AEB	1.29	7.1	322
51	2B0T	1.75	7.3	738
52	2BG1	1.9	7.0	494
53	2BWR	1.5	6.5	401
54	2C0H	1.6	6.5	353
55	2CJP	1.95	7.5	328
56	2DSJ	1.80	7.3	423
57	2EX2	1.55	6.5	458
58	2FAO	1.50	6.5	309
59	2GGC	1.00	7.0	263
60	2H2Z	1.60	6.0	306
61	2H3L	1.00	7.5	103
62	2HDW	2.00	6.5	367
63	2HKJ	2.00	7.5	469
64	2HLY	1.60	6.0	207
65	2O2K	1.60	7.5	355
66	2OIT	1.65	6.5	434
67	2PEF	1.60	7.5	373
68	2PKF	1.50	6.5	334
69	2PUH	1.82	7.0	286
70	2PVQ	1.80	7.0	201
71	2PYW	1.90	6.5	420
72	2QMM	1.85	7.5	197
73	2V3V	1.99	6.5	723
74	2VFR	1.10	6.5	422

75	3A3P	1.90	6.5	329
76	3B40	2.00	7.0	417
77	3B64	1.03	7.5	112
78	3BB7	1.50	6.5	321
79	3BF7	1.10	7.0	255
80	3BY9	1.70	7.5	260
81	3CFZ	1.70	7.0	292
82	3CNQ	1.71	6.5	292
83	3CSS	1.70	7.5	267
84	3CUZ	1.04	7.5	532
85	3DRF	1.30	7.0	590
86	3DSK	1.55	7.5	495
87	3E7H	1.70	6.5	103
88	3EA1	1.75	6.5	298
89	3EH1	1.80	7.5	751
90	3EHR	1.95	6.5	222
91	3F1L	0.95	7.5	252
92	3FOT	1.75	7.5	519
93	3GGQ	2.00	7.5	149
94	3H2G	1.86	6.7	397
95	3HUH	1.50	7.5	152
96	3HXL	1.90	7.5	446
97	3IFE	1.55	6.5	434

Table 2.13. Results of super long loop prediction (14-20 residues).

PDB ID	Loop start	Len-gth	VSGB 2.0			VSGB 1.0		
			EGAP (kcal/mol)	Backbone RMSD (Å)	Side Chain RMSD (Å)	EGAP (kcal/mol)	Backbone RMSD (Å)	Side Chain RMSD (Å)
1E6U	A274	14	0	0.27	0.85	-7	1.94	2.93
1JP4[a,b]	A153	14	-4	0.31	0.67	-43	7.26	6.32
1N0Q	A24	14	-12	0.51	1.55	-4	0.22	1.76
1N0Q	A57	14	-10	0.50	2.29	-13	0.52	2.34
1O97	D156	14	-8	0.68	1.93	-13	0.82	1.55
1OCK	A209	14	-18	1.13	1.80	-22	0.92	1.86
1P3C	A112	14	-18	0.31	1.28	-7	0.32	0.94
1P3D	A402	14	0	0.35	0.40	6	1.74	4.99
1R6X	A72	14	2	0.25	0.41	1	0.30	0.48
1RDQ	E273	14	-9	1.34	2.74	-11	1.50	2.11

1RV9	A225	14	3	0.28	1.92	9	0.26	2.20
1VYR	A193	14	-6	0.51	1.76	4	0.58	1.57
1VYR	A235	14	-7	0.93	1.98	6	1.17	3.06
1XU1	A221	14	-4	0.69	2.12	4	0.47	1.58
1ZEQ	X53	14	-18	0.21	2.10	-15	0.27	2.01
2BWR	A269	14	-22	0.31	1.70	-3	0.44	2.35
2BWR	B158	14	-2	0.41	1.85	35	4.56	7.40
2C0H[b]	A40	14	0	0.55	1.56	108	5.96	8.85
2EX2	A139	14	-18	0.32	1.50	-8	0.21	1.47
2GGC	A79	14	-8	0.28	1.70	-8	0.24	1.95
2H3L	A1360	14	-12	0.36	2.10	1	0.28	1.84
2O2K	A1221	14	-1	0.29	1.47	-29	1.07	2.36
2PVQ	A139	14	-13	0.57	1.31	6	0.66	1.62
2VFR	A325	14	-3	0.19	0.33	-14	0.34	1.43
3B40	A389	14	1	0.33	0.89	48	1.28	2.07
3B64	A44	14	-1	0.24	1.04	-25	0.65	2.09
3BY9	A177	14	2	0.47	1.66	-24	1.13	2.70
3BY9	A205	14	-6	0.32	0.81	-11	0.28	0.81
3CFZ	A125	14	-8	0.54	1.99	1	0.68	1.92
3CNQ	S50	14	-2	0.63	1.88	-9	1.03	2.28
3CSS	A163	14	0	0.19	0.27	0	0.21	0.27
3DRF[b]	A550	14	-10	1.04	2.51	1	1.85	3.38
3E7H	A67	14	-3	1.27	3.41	-6	1.78	2.85
3EHR	A95	14	-4	0.51	3.40	7	0.94	1.78
3FOT	A164	14	-14	0.34	1.41	-14	0.30	1.40
3HXL	A277	14	0	0.77	3.74	28	0.79	3.92
1AH7	A157	15	-13	0.55	1.75	-1	0.32	1.11
1BHE	A121	15	-3	0.76	2.52	-1	0.42	1.88
1H4A	X19	15	-1	0.31	1.22	-4	0.28	1.26
1JU3	A486	15	0	1.09	2.51	5	0.35	1.64
1QAZ	A298	15	-29	0.99	3.00	-7	1.68	3.72
1QQF	A1112	15	-2	0.53	0.69	-8	0.31	2.03
1RA0	A283	15	7	0.30	2.16	-39	2.78	5.46
1RA0	A361	15	-5	0.52	1.66	-4	0.39	2.15
1RYO	A172	15	-1	1.17	1.60	-5	0.88	1.64
1S95	A477	15	-15	0.46	1.86	-5	0.61	2.09
1WB4	A1033	15	-1	0.15	0.47	1	0.21	0.80
1WUI	L454	15	2	0.97	2.96	-12	1.81	3.76
1Y12	A10	15	-2	0.62	2.81	-2	0.36	2.10
1ZHX	A392	15	4	0.35	1.12	67	7.10	8.21

2AEB	B156	15	5	1.32	2.69	25	2.55	5.23
2B0T	A701	15	1	0.94	1.63	3	1.23	3.74
2CJP	A58	15	-9	0.43	0.77	12	0.46	0.98
2DSJ	A354	15	-7	0.57	1.89	-5	0.51	1.85
2H3L	A1339	15	-10	0.35	2.02	-10	1.10	1.88
2O2K	A1220	15	13	1.53	3.28	-29	1.36	3.38
2OIT	A290	15	7	0.80	2.32	6	0.54	2.30
2PKF	A26	15	0	0.65	1.59	-8	2.34	5.06
2V3V	A382	15	-6	0.41	1.78	4	0.35	1.94
3A3P	A286	15	4	0.22	1.71	5	0.18	0.99
3A64	A350	15	-7	0.20	1.22	3	2.55	3.34
3BB7	A231	15	-9	0.33	2.06	-19	6.26	7.60
3BF7[b]	A49	15	3	0.58	1.58	72	5.66	9.28
3CSS	A95	15	-17	0.52	1.39	-24	2.36	3.79
3EA1	A136	15	17	0.76	1.79	7	0.49	1.08
3F1L	A99	15	-5	0.50	1.44	-16	1.05	1.76
1C1K	A31	16	-8	1.71	2.67	12	0.66	1.67
1DJ0	B19	16	6	1.66	4.24	72	7.08	13.52
1GPI	A308	16	-11	0.52	1.04	-5	0.33	0.68
1UG6	A340	16	4	0.37	1.08	4	0.43	1.03
1WHI	A88	16	-13	0.47	2.34	-17	0.66	1.79
1WM3	A67	16	-10	0.23	1.24	13	0.32	0.52
1ZHV	A20	16	-6	0.30	1.03	-10	0.64	1.05
2BG1	A708	16	-3	0.40	0.92	-4	2.15	2.76
2GGC	A184	16	-12	1.17	2.33	-89	1.08	3.84
2HKJ	A418	16	-16	0.36	1.11	-42	0.42	1.59
2PKF	B25	16	-42	0.95	2.89	-93	0.93	3.38
2PUH	A70	16	-5	0.30	0.96	-14	1.52	3.25
2PYW	A321	16	-1	0.99	3.45	-37	2.57	6.1
3IFE[b]	A14	16	-13	0.32	0.57	-55	1.18	3.68
1KWG	A314	17	0	0.57	1.73	-41	1.93	2.92
1QLW	A145	17	-8	0.35	2.11	0	0.41	1.29
1VJU	A277	17	4	0.38	1.82	-2	0.63	1.55
2FAO	A814	17	3	0.86	1.84	21	0.50	1.27
2HDW	A131	17	10	0.60	1.63	-126	2.22	5.96
2PEF	A191	17	2	1.32	2.97	44	1.92	3.55
3A3P	A262	17	16	0.21	0.96	-250	0.52	2.79
3H2G[b]	A124	17	-24	1.00	1.50	56	4.07	6.75
3HUH[b]	A71	17	-6	0.29	1.98	-39	8.56	12.13
1GQ6	A239	18	-1	0.29	0.62	-133	1.47	3.45

1NC5	A121	18	16	0.26	0.91	129	9.78	13.27
1J83	A1073	18	11	0.27	0.96	47	4.86	5.46
1R6D	A132	18	13	0.35	1.54	11	0.67	1.28
1RQW	A69	18	6	1.29	3.93	1	0.43	3.24
1UW4	B899	18	5	1.97	1.79	-102	11.27	11.47
1WRJ	A53	18	-10	0.49	1.27	-30	1.55	2.79
1XFK	A258	18	-7	0.56	1.33	8	1.42	1.92
2AEB	A153	18	-5	1.15	2.71	-	-	-
2H2Z[b]	A130	18	-3	1.49	2.86	-3	1.48	2.07
2QMM[b]	B103	18	5	0.79	2.15	34	8.95	9.18
3CUZ	A384	18	4	0.82	2.00	-22	3.45	5.94
3DSK	A425	18	12	0.39	1.03	26	5.78	7.50
3EA1	A164	18	10	0.65	1.45	108	3.70	4.23
3EH1[b]	A813	18	11	1.60	2.37	-22	2.84	4.58
3GGQ	A550	18	8	0.37	1.54	-26	5.12	7.46
1NC5	A193	19	-12	1.79	3.81	1.06	3.66	4.96
1ODM[b]	A119	19	16	1.94	3.39	5.59	1.31	2.50
1Q0R [b]	A58	19	7	1.51	5.07	-3.99	3.87	5.37
1QL0	A150	19	-6	1.02	1.72	-7.68	1.29	2.36
1V8H[b]	A27	19	11	1.74	5.42	15.95	7.59	8.64
2HLY	A126	19	10	1.60	3.13	-1.08	0.48	1.54
2O2K	A947	19	3	0.25	1.70	0.73	0.33	1.72
1QL0	A47	20	-2	1.29	2.45	-0.73	0.67	1.93
1YW5[b]	A76	20	4	1.80	3.18	-4.51	2.51	3.45
2AEB[b]	A50	20	-16	1.68	3.02	-9.02	1.12	2.75

a. Explicit treatment with key water molecules between crystal copies.
b. Extended sampling to "Fix10" stage.

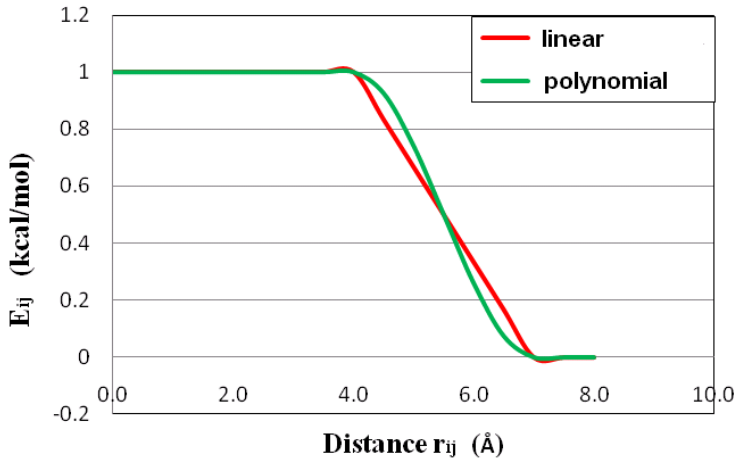


Figure 2.11. Comparison of the linear function and the polynomial of hydrophobic term (for one pair of interacting nonpolar atoms).

Chapter 3. Predictions of P450-mediated Drug Metabolism

3.1. Introduction

It is crucial to understand how potential drugs are metabolized in the body, because human metabolism has profound impacts on the bioactivity and the safety profiles of drug candidates. On one hand, metabolism can convert these compounds into their active forms, which interact with the therapeutic targets; on the other hand, metabolism eliminates the compounds by converting them into inactive excretable metabolites. Sometimes the metabolic modifications also lead to toxicity, which can cause unexpected failures in the later phases of drug development. Furthermore, the metabolic behavior of drug compounds is also highly related to other critical issues such as food-drug interactions, drug-drug interactions, and personalized medication.[103-105] Given the enormous impact of metabolism on drug bioavailability and toxicity, it is important to determine metabolites in the early stage of the drug discovery process. However, to obtain such information experimentally is often a very lengthy and expensive process. Therefore, it would be extremely useful if one could use computational methods to predict the metabolic decomposition of drug candidates.

Since cytochrome P450 enzymes (CYP) are involved in a large majority of drug metabolism pathways, many computational studies have been published attempting to predict P450-mediated metabolism using a variety of methods and models. For a recent review see the work of Afzelius *et al.*[106] These previous studies mainly focused on the important P450 isoforms 2D6, 2C9 and 3A4 aiming to predict the primary metabolites of drug compounds. Several ligand-based methods have been developed during the past decade, making predictions based on hydrogen abstraction energies estimated with semiempirical quantum mechanics[107] or DFT methods[108]. Although such ligand-based methods are very fast, it is often necessary to consider

the interaction between the enzyme and the substrate in order to reach high accuracy (for example, >80% agreement with experiments) in the predictions. It is possible to include a limited amount of enzyme specific information by making descriptors of ligand-based models dependent on the nature of the enzyme.[41,109,110] Such approaches have been successfully implemented in software packages such as MetaSite and some were reported to recover up to 86% of the experimental observations.[111] On the other hand, molecular dynamics (MD) or induced-fit docking simulations in combination with transition state calculations at the QM/MM or semiempirical quantum level were used to predict metabolites for a few ligands.[112,113] Other promising methods based on molecular docking have been implemented as well,[114-118] which determine the predictions using a reactivity model and/or distance cutoffs from the reactive iron center.

Traditional empirical ligand-based approaches to the prediction of P450 SOMs rely primarily on implicit estimation of intrinsic site reactivity to the Compound I oxo species, coupled with a heuristic attempt to take into account the ability of the ligand to bind to the P450 active site. While such methods can yield some discrimination of true positives from false positives when a sufficiently large training set is employed,[41,107,119] the precision of the approach is fundamentally limited, as the treatment of protein-ligand binding is highly approximate. Methods such as MetaSite[111] provide some incorporation of P450 structural information, but employ a much smaller training set and fewer empirical parameters; the overall results appear to actually be less accurate than a ligand-based approach employing an extensive data set. The problem is again that the MetaSite algorithm for modeling the reactive protein-ligand complex does not rigorously evaluate the binding energy, or perform a thorough conformational search, severely limiting the predictive capability that can be attained.

The method described in the present work (IDSite) represents a qualitatively different approach from those discussed above, as well as from other efforts in the literature.[41,108,110,111] Firstly, the goal is to actually generate an accurate structure for the protein-ligand complex that enables reactivity at a specified site; this requires construction of a good approximation to a transition state structure for both aliphatic and aromatic sites of reaction. Secondly, the relative binding affinity, as compared to alternative structures for both the site in question and for other sites, has to be computed with a respectable degree of precision, on the order of a few kcal/mol. Finally, The relative intrinsic barrier height of the reaction (combined with the relative binding affinity to produce an overall relative barrier , as compared to other possible reactions of the molecule, must be estimated, to within ~1 kcal/mol. These are extraordinarily daunting tasks, given that the P450 isoforms present large, complex active site regions with substantial capability for induced-fit conformational changes, a necessary condition for them to accommodate the wide range of exogenous ligands with which they need to interact to perform their biological function.

The algorithms in IDSite employ a novel model for the total energy of the protein-ligand complex, which has recently been show to provide remarkably accurate predictions for side chains and loops[120], and a sophisticated algorithm for generating converged induced-fit structures which combines docking, conformational search, and hybrid Monte Carlo (MC) methods based on MD trajectories. The algorithm enables a hierarchical search which addresses the various length scales of the problem, including the small correlated motions provided by the MD trajectories which we have found are absolutely necessary to produce useful rank ordering of structures, particularly for larger ligands. Constraints are employed in conjunction with these simulation algorithms to enforce appropriate transition state structures. The energy model

enables the targets of a few kcal/mol accurate in relative binding affinity to be reached. Finally, a quantum chemically based model is employed to calculate relative intrinsic reactivities, and again is shown below to yield outstanding performance. Based on such high level of success, it is documented below in predicting true positive SOMs vs. false positives.

To our knowledge, these results represent the first reliable and accurate computation of binding poses and transition states for a wide range of drug-like molecules interacting with an important human P450 isoform. There are a few previous papers in which structures are generated via QM/MM calculations,[113,121,122] however these typically address a very small number of ligands (usually one), the ligands are typically simpler and smaller than those treated here, and the sampling algorithms are much less extensive. We believe that these structures can be very useful in practical drug design applications, in situations where modification of P450 metabolic properties for candidates in later stages of lead optimization is required. The availability of an atomic level three dimensional structure, as well as the ability to predict the structural and energetic effects of chemical modification of the molecule, provides a new tool for chemists to rationally engineer desirable metabolic properties into clinical candidates. Extension of our methods to other P450 isoforms such as 2C9, 3A4 and 1A2, which is currently in progress, will enhance the utility of our approach for this important application.

3.2. Methods and Materials

Overview of IDSite Methodology

IDSite combines the docking program Glide[123] and the protein structure modeling program PLOP (Protein Local Optimization Program, available as the protein refinement module in the protein modeling package Prime of Schrödinger, Inc.[124]), to model induced fit effects and to predict sites of metabolism. IDSite consists of three hierarchical sampling stages and one final

scoring stage (Figure 3.1). It begins with flexible Glide docking calculations, which place the ligand into the active site. Following the docking stage, two refinement stages in PLOP are carried out to refine the protein side-chain and ligand orientations. At the end of each sampling stage, the generated/refined poses are screened based on their structures and energies, and clustered according to the similarity of the ligand conformation. Finally, the refined lowest energy poses are used to predict the sites of metabolism based on a physical score, which is dependent on the energies of the poses as well as the intrinsic chemical reactivities of the potential sites of metabolism.

IDSite is able to use knowledge about specific conserved interactions to perform efficient sampling and accelerate the calculations. For example in the case of CYP2D6, a typical substrate always contains a basic center (e.g. an amine nitrogen) that binds to one of the two acidic residues, Glu216 or Asp301. IDSite constrains such salt bridges to reduce the sampling cost associated with the docking and refinement stages. Filters are applied during the screening at the end of each stage in IDSite to reduce the number of poses passed to further refinement or evaluation (Table 3.1). The following is a detailed description of each stage of IDSite.

We have constructed our sampling and scoring algorithms with the intention of approximating the correct transition state structure of the protein-ligand complex and associated activation energy which would lead to reactivity of the target atom of the ligand. There are two components of the problem: finding the transition state in reasonable CPU time (a daunting task for a large, complex ligand when induced fit effects are important), and estimating the free energy of activation associated with the transition state. The VSGB 2.0 energy function, with constraints to enforce a suitable geometry for the reaction to take place (and some other constraints as well to facilitate sampling, as described in the text below), is minimized to

generate these structures for the various possible candidate reactant heavy atoms. We use the classical force field and solvation model to produce a “reactant” structure which is optimally positioned for the targeted chemical reactivity. The activation barrier from a precomputed quantum chemical fragment calculation, as described in the following text, is then added to the VSGB 2.0 energy to estimate the relative energy barrier for converting such a structure into products. This is an approximation to a more rigorous approach such as using QM/MM methods to generate the reactant, transition state, and product structures. Note that it is only important that relative free energy of the various potential sites of reaction are calculated with reasonable accuracy, as the most reactive (lowest activation free energy) site is always used as a reference point (i.e. the energy function for this site is subtracted from the energy function for the candidate site) in our assessment of the metabolic contribution of each site. Finally, in applying the above protocol, the VSGB 2.0 energy must be calculated using a structure with the constraints in place, otherwise the structure would minimize to something that is not a suitable starting point for reaction. The constraints introduce some strain energy into the structure, but this strain energy is an appropriate component of the activation free energy as it does cost energy to create a suitable reactive structure.

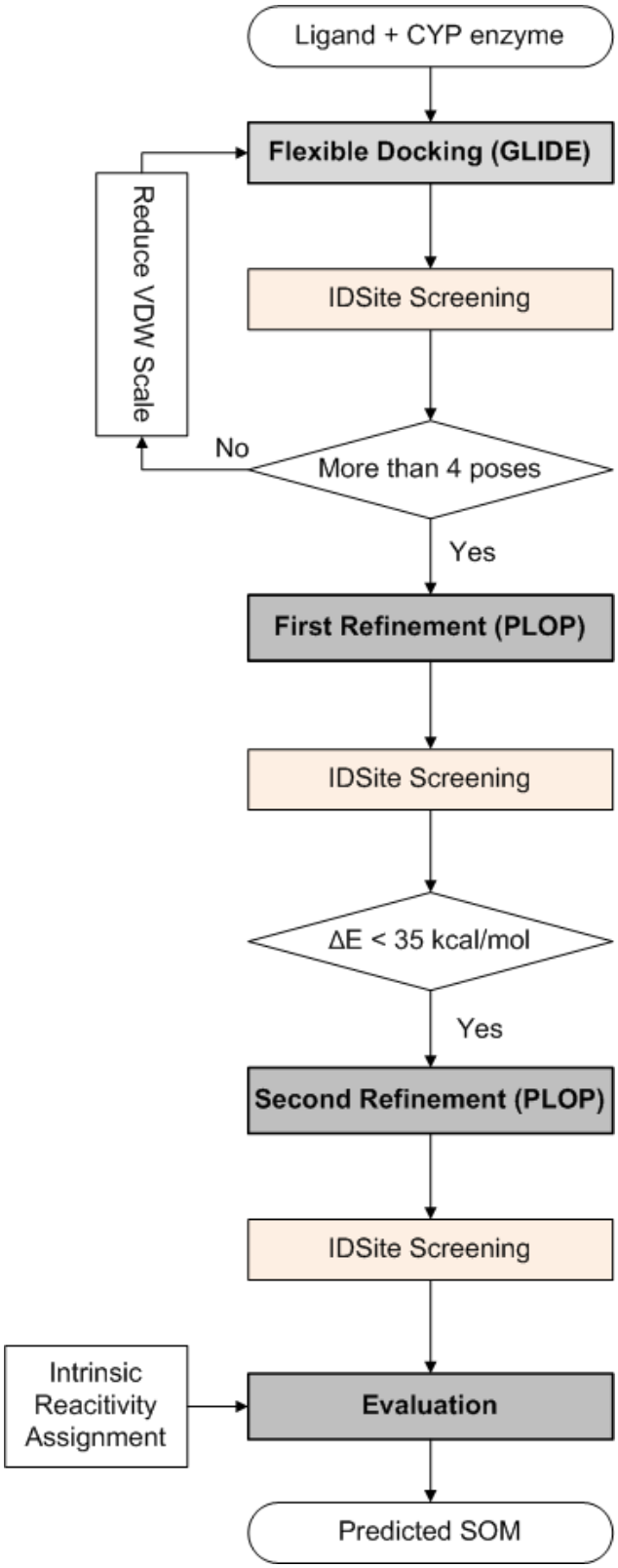


Figure 3.1. IDSite workflow.

Table 3.1. IDSite filters in the screening for CYP2D6.

Stage	Filters applied in the screening at the end of the stage
Glide Docking	Poses that fulfill any of the criteria below are removed: 1) The distance of the basic nitrogen to the ferryl oxygen is less than 5.0 Å; 2) The distance of the basic nitrogen to the negative charged oxygen (in Glu216 or Asp301) is greater than 5.5 Å; 3) More than 2 heavy atoms from the ligands are further than 16.0 Å away from the heme iron;
PLOP Refinement 1	4) More than 1 heavy atom from the ligand are closer than 1.0 Å to the receptor; 5) More than 6 heavy atoms from the ligand are closer than 1.8 Å to the receptor; 6) No heavy atom in the ligand is within 5.0 Å to the heme iron. For PLOP Refinement 1: All the poses are ranked with PLOP energies. Poses with energy higher than 35 kcal/mol compared to the lowest energy pose are removed.
PLOP Refinement 2	Poses that fulfill any of the criteria below are removed: 1) the distance between the constrained atom and the ferryl oxygen is outside the optimal range which is from 1.65 to 2.60 Å for sp^3 atoms and from 1.60 to 2.08 Å for sp^2 atoms; 2) the distance of the basic nitrogen to the ferryl oxygen is less than 4.8 Å; 3) the distance of any polar atom to the ferryl oxygen is less than 3.2 Å; 4) the distance of the constrained salt bridge (between the basic nitrogen and the oxygen from Glu216 or Asp301) is greater than 3.6 Å; the angle of the salt bridge (N-H-O) is less than 140 degree; 5) more than 2 heavy atoms from the ligands are either further than 14.5 Å or closer than 1.6 Å from the heme iron; 6) The pose has at least 1 distorted cyclohexane ring.

Glide Docking

Starting from the ligand and the protein receptor structures, IDSite carries out flexible ligand docking with Glide.[18,19] The flexible ligand docking protocol generates a large number of ligand conformations that are then docked into the rigid receptor. The first step in Glide docking is to define the binding box and calculate the receptor grid. As in Glide, in IDSite the binding site is defined as a box centered at the center of selected residues or a ligand (if the structure contains a ligand). Because we start from the apo structure of CYP2D6 (PDBID: 2F9Q. See below for details about the protein preparation), the center of the binding box is selected as the centroid of the residues Glu216, Asp301, Thr309, and Phe483. The box dimension on each side is set to 10 Å for the inner box and 20 Å for the outer box. After the grid generation, IDSite samples the conformations of freely rotatable bonds and rings with Glide Standard Precision (SP). In order to increase sampling, IDSite uses reduced Van der Waals (VDW) radii and skips the default filtering with a rough score within Glide (also referred to as expanded sampling). Similar poses are clustered according to their RMSD (cutoff 2.0 Å). Finally, a post-docking minimization is performed and the top 60 minimized poses according to the Glide SP score are retained. These poses are then screened to remove the poses with obvious steric clashes, with too many atoms outside the inner binding box, or without atoms close to the heme iron (Table 3.1). The remaining poses are then passed to the first refinement stage.

IDSite uses reduced VDW radii for nonpolar atoms both in the protein receptor and the ligand, so that slight steric clashes are tolerated during the docking stage. For the protein receptor the VDW scaling factor is fixed at 0.40, while for the ligand, the scaling factor starting from 0.80 is adaptively adjusted until at least 4 valid poses are found. With highly flexible ligands and relatively high scaling factors, Glide often finds only a handful of valid poses, and even fewer survive after IDSite screening. However, if the scaling factor is set too low, the docked poses

may contain too many serious steric clashes, which can cause problems in the subsequent minimization. If IDSite fails to find enough valid poses, the scaling factor is adjusted and the number of poses to pass the initial docking phase in Glide is increased accordingly to augment sampling.

Since a typical CYP2D6 substrate forms a highly conserved salt bridge with either Glu216 or Asp301,[125] IDSite employs this conserved interaction to reduce the sampling cost of the CYP2D6-docking in the following way: IDSite adds a positional constraint to ensure that the generated poses fulfill at least part of the preferred conserved interactions. The positional constraint defines a spherical region in the receptor that is within 4.0 Å of the center of the Glu216, Asp301, and Ser304 residues (Figure 3.2). It is required that during docking and post-docking minimization each pose should maintain at least one hydrogen-bond donor inside the spherical region. If the ligand contains other hydrogen-bond donors except for the basic nitrogen, the constrained docking is likely to generate poses that form hydrogen bonds instead of the salt bridge to Glu216 or Asp301. However, IDSite is able to distinguish these poses and filter them via an additional salt bridge filter in the pose screening (Table 3.1), so that only the poses with a stable salt bridge are allowed to pass to the refinement stage.

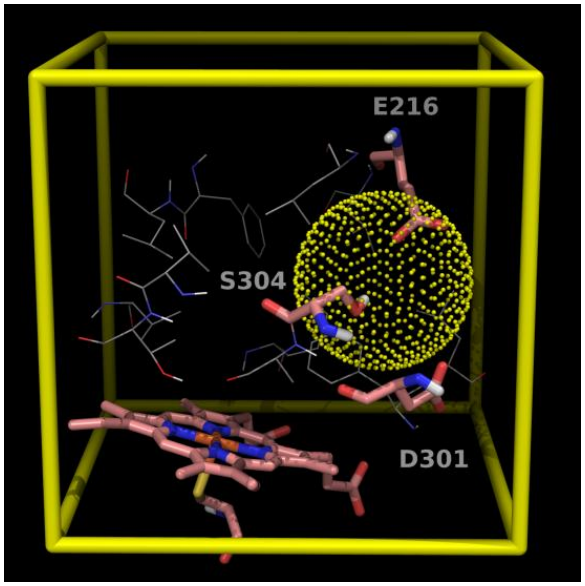


Figure 3.2. Definition of the binding box (yellow cube) and the positional constraint (yellow dotted sphere) in IDSite for CYP2D6.

PLOP Refinements

The refinement of the docked poses includes multiple, parallel Monte Carlo Minimization (MCM) simulations in PLOP. For each pose from the previous stage (the docking or first refinement stage), IDSite finds all the heavy atoms in the ligand close to the heme iron. For each of these atoms, distance and angular harmonic constraints are applied in order to force sampling of the conformations that potentially lead to metabolism. The optimal distances and angles of the constraints were obtained from hydroxylation transition state geometries with a heme model system at the B3LYP/LACVP* level using Jaguar.[126] The detailed nature of the employed constraints is shown for both sp^3 and sp^2 type carbons in Figure 3.3 and 3.4. The constraints are then employed in the minimizations step but were not included in the energy used for the acceptance step of the MCM simulations. PLOP uses the overlap factor (the ratio of distance between two atoms centers to the sum of their atomic radii) to quickly reject randomized structures with serious steric clashes (defined as the overlap factor being lower than a specific

cutoff). PLOP repeats the random attempts until a structure with tolerable clashes is generated, after which a constrained minimization using the truncated Newton method is performed. The acceptance or rejection of the minimized structure is decided by the Metropolis criteria based on the energy calculated in the VSGB 2.0 model. (Performing the minimization step before testing the acceptance criteria violates detail balance, but this is not an issue as we are interested only in low energy structures and not the population/ensemble distribution.) The simulations run until a certain number of accepted structures are collected.

In order to sample the various degrees of freedom in the conformational space, IDSite employs three types of randomized moves in the MCM simulations: side-chain rotation, rigid body translation/rotation, and hybrid moves.

Side-chain moves: By varying the dihedral angles of the rotatable bonds, IDSite uses side chain MC moves in PLOP to sample the selected side-chain conformations of the protein and of the ligand. Up to three close residues (C_β distance within 6 Å) are allowed to rotate collectively, but the moves of the protein residues and those of the ligand are separated. In each attempted movement, the conformations of the selected side chains (from the protein/ligand) are either changed by random perturbations or assigned by the randomly selected rotamers from a library. For an attempt with a random perturbation, the displacement of each dihedral angle is the sum of a large rotation (N times 60 degrees with N as a random integer between 0 and 5) and a random perturbation from 0 to 30 degrees. For a rotamer library attempt, a side-chain conformation is updated with a random rotamer from a high resolution side-chain library for protein residues[79], and from a homogeneous library at 10 degree resolution for the ligand. If a structure with tolerable overlaps is generated in an attempt, it is minimized and sent to subsequent stages for

judgment of acceptance. Each side-chain move takes less than 15 seconds and is the fastest among all the three move types.

Rigid body moves: Rigid body moves are used to sample the translational and rotational space of the ligand. Multiple attempts with reduced VDW radii are applied, as it is quite common to fail in searching for a clash-free conformation in a single rigid body moving attempt (especially when the ligand is large and flexible and the binding pocket is relatively small). Each rigid body move includes 1000 attempts, and each attempt performs a translation along a random vector and a rotation around a random axis, with less than 0.5 Å and 60 degree displacement, respectively. In addition, the VDW radii are reduced (scaling factor 0.8) to soften the Lennard-Jones potential, so that mild steric clashes are allowed, which are likely to be resolved by the subsequent minimization. The rigid body move usually takes 20 to 40 seconds per move.

Hybrid Monte Carlo moves: The hybrid Monte Carlo (HMC)[127] move in PLOP performs simultaneous sampling for the selected residues in the protein side chains and backbone as well as the ligand. Each HMC move performs a 5 picosecond, constant energy molecular dynamic (MD) simulation (starting at 900K) on all the atoms in the selected residues. Taking up to 15 minutes per move, the HMC is the most expensive among all three types of moves in PLOP.

Considering the different costs for the three types of moves, the frequency of deployment of each move type in the various refinement stages is adjustable according to the sampling requirements. Two stages of refinement with different combinations of moves and constraints are carried out in the hierarchical sampling. Using more HMC moves, the first refinement stage applies loose distance constraints between an atom in question (from the ligand) and the ferryl

oxygen. It is designed to "pull" the close atom (identified from the docking poses) towards the heme iron, to estimate the likelihood that the atom can approach the iron and react with the ferryl oxygen. When an atom in the ligand is forced to be proximate to the ferryl oxygen under the constraints, the rest of the ligand and the surrounding protein residues have to adjust their conformations accordingly. The adjustments for some poses are easy while for some others are difficult, depending upon the specific geometrical issues and energetics of the protein-ligand interactions for the trajectory connecting particular starting and target poses. Resulting poses with steric clashes or distorted structures can be identified by their high energies and discarded in the IDSite energy and structure screening (Table 3.1). The low energy poses after screening, mostly with favorable interactions between the protein and the ligand, are passed to the second refinement stage. Mainly focusing on side-chain sampling, the second refinement stage applies tight constraints that force the structure to form special conformations similar to that of the transition states obtained from DFT calculations of model systems. The second refinement stage is used to further refine the poses and distinguish the potential of each atom in question to be oxidized. The comparison of the settings for these two refinement stages with PLOP are shown in Table 3.2, while the constraints are illustrated in Figures 3.3 and 3.4. There are approximately 39 protein residues, identified to be important for ligand binding by mutagenesis experiments or are adjacent to these key residues that are sampled during the refinement stages. At the end of each refinement stage, all the poses sampled in that stage are screened and clustered for further refinement or evaluation (Table 3.1).

For CYP2D6, harmonic constraints are also applied to force the basic nitrogen to interact with the acidic residues, Glu216 and Asp301 (Figure 3.5 and 3.6), as they are believed to play important roles in substrate binding to CYP2D6 from mutagenesis experiments.[128,129]

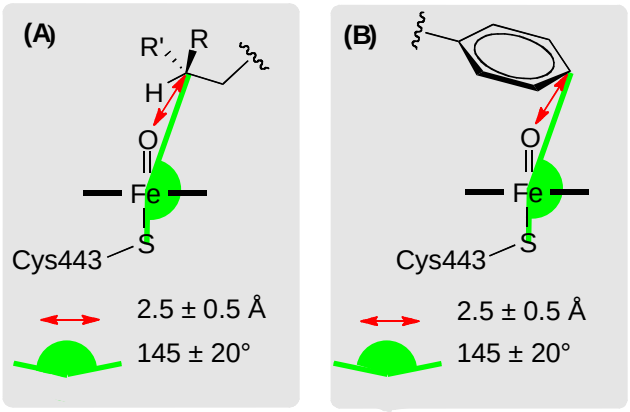


Figure 3.3. Constraints applied to the heme region in the first refinement stage. The ferryl oxygen is a "dummy" atom (1.6 Å above the heme iron), only used to define the constraints in the IDSite calculations. **(A)** Constraints for sp^3 carbons. **(B)** Constraints for sp^2 carbons.

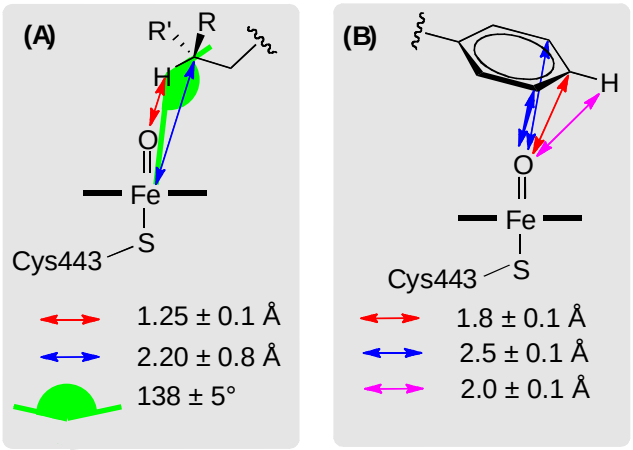


Figure 3.4. Constraints applied to the heme region in the second refinement stage. The ferryl oxygen is a "dummy" atom (1.6 Å above the heme iron), only used to define the constraints in the IDSite calculations. **(A)** Constraints for sp^3 carbons. **(B)** Constraints for sp^2 carbons.

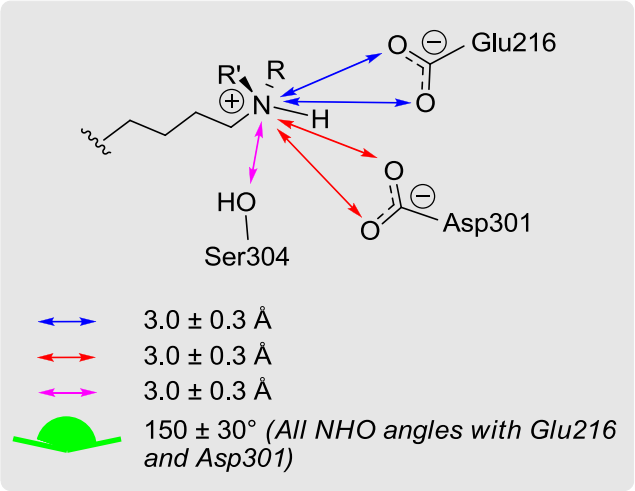


Figure 3.5. Constraints applied to the salt bridge region of CYP2D6 in the first refinement stage.

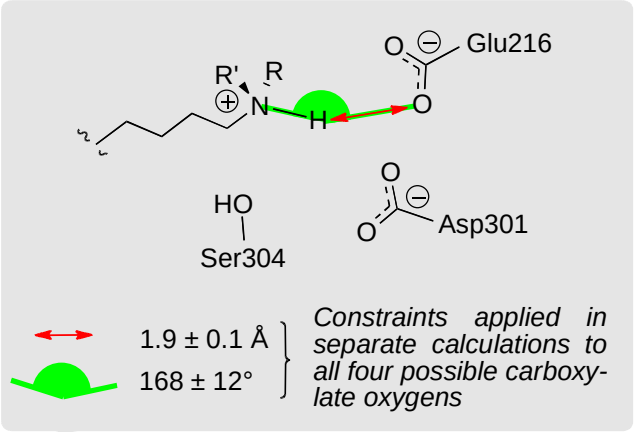


Figure 3.6. Constraints applied to the salt bridge region of CYP2D6 in the second refinement stage.

Table 3.2. Comparison of settings in the first and second refinement stages.

	PLOP Refinement 1	**PLOP Refinement 2**
Number of residues to sample (including the ligand)	12	40
Number of accepted structures for each job	Maximum of 8 times the number of rotatable bonds, 24	Maximum of 20 times the number of rotatable bonds, 60
Types and probabilities of MCM moves	Side chain: 0.50 Rigid body: 0.10 Hybrid: 0.40	Side chain: 0.70 Rigid body: 0.10 Hybrid: 0.20

Evaluation

Herein, we present two scoring models to evaluate the potential sites of metabolism and to determine the predictions. Our first scoring model (referred to as Physical IDSite) is based on the following assumptions: (1) For hydroxylation of an aliphatic chain carbon, the P450-hydrogen abstraction step is rate determining.[130,131] (2) For hydroxylation of aromatic rings, the electrophilic attack of Compound I on the aromatic ring is rate determining.[130,131](3) All reaction intermediates before the rate determining step are in equilibrium.[132] Given these assumptions, the relative rates of product formation depend only on the relative transition state free energies of the rate determining (RD) transition states ($\Delta G^{\neq}$) according to the Curtin-Hammett principle. These can then simply be written as

$$\Delta G^{\neq} = \Delta G_{bind} + \Delta G^{\neq}_{RD\text{-}step} \qquad \text{(Eq. 3.1)}$$

where ΔG_{bind} is the binding free energy of the substrate into the reactive conformation in the P450 active site and $\Delta G^{\neq}_{RD\text{-}step}$ is the activation barrier of the RD-step.

In the present application of IDSite, we attempt to calculate only relative, as opposed to absolute, site reactivity for a given ligand. Absolute site reactivity for the ligand can typically be obtained via inexpensive experiments. However, detailed metabolic chemistry is often more difficult to determine, and an accurate three dimensional structure leading to reactions at each metabolic site is not available given the severe challenge of obtaining a crystal structure of a P450 isozyme with the ligand bound in the reactive conformation. Prediction of the most highly reactive site, followed by identification of all sites with relative reactivities sufficiently large to be experimentally detected along the dominant metabolic pathway, coupled to structural prediction for each relevant reactive geometry, complements current experimental practice and facilitates compound modification in situations where P450 metabolism needs to be altered to confer improved metabolic properties on a candidate drug molecule.

In the Physical IDSite model, the relative binding energies of various docked poses are calculated from the PLOP VSBG 2.0 energies of these poses, while the barriers for the RD-steps are estimated from

the corresponding activation barriers of model compounds with a methoxy radical (calculated at the DFT level). E_{pose} in Eq. 3.3, calculated in PLOP, estimates the protein-ligand interactions when a potential site is forced to approach the catalytic center in a certain pose with a transition state-like conformation. Based on the linear correlation (Figure 3.7) between the methoxy radical activation barriers and the corresponding activation barriers with the heme system, we approximated the real activation barrier for each potential site of metabolism from the intrinsic reactivity calculated with the methoxy radical model according to Eq. 3.2.

$$IR(heme) = 1.117 \times IR(methoxy\ radical) + constant \qquad \text{(Eq. 3.2)}$$

With the constant from Eq. 3.2 ignored, the relative $\Delta G^{\neq}$ for each potential site (approximated as the score E) is then calculated as the Boltzmann weighted average over the energies of all contributing poses, where angle brackets represent the Boltzmann averages (Eq. 3.3). A term describing the configurational entropy of equivalent hydrogen atoms at 298 K, proportional to the logarithm of the number of symmetrically equivalent hydrogen atoms, was also included. The $\Delta G^{\neq}$ values for all symmetrically equivalent sites were set to the lowest $\Delta G^{\neq}$ of the sites.

$$E = \langle 1.117 \times IR(methoxy\ radical) + E_{pose} \rangle - kT\ln N_H \qquad \text{(Eq. 3.3)}$$

Since (as a rule of thumb) it is difficult to observe a minor metabolite experimentally if it is formed in less than ca. 0.1% yield (which corresponds to ca. 4.75 kcal/mol increase in relative $\Delta G^{\neq}$ compared to the free energy of the most favored product), we used 4.75 kcal/mol as a cutoff for the prediction; with Physical IDSite, any potential sites of metabolism having a relative $\Delta G^{\neq}$ lower than 4.75 kcal/mol is predicted to be a site of metabolism.

The second scoring model represents an empirically optimized version of the physical model described above with the following changes: (1) The PLOP energy (E_{pose}) is not used directly, but rescaled with two parameters as described below, which are fitted to a training set of 36 compounds. (2) Instead of obtaining the scaling coefficient for the methoxy radical intrinsic reactivities from the

correlation in Eq. 3.7, we fit it to the training set of 36 compounds. Note that the fitted value for the latter of 1.071 (Eq. 3.4) is very similar to the value obtained by correlating the DFT-activation energies (1.117), which further highlights the physical nature of this parameter. (3) The final selection criteria for predictions (score cutoff) were fit to the training set as well. All four fitted parameters were obtained from a fitting algorithm by maximizing the number of true positives over the sum of the numbers of false positives and false negatives.

$$E = <1.071 \times IR(\text{methoxy radical}) + E_{score}> - kT\ln N_H \qquad \text{(Eq. 3.4)}$$

As introduced above, instead of directly using the PLOP energy (E_{pose}), Eq. 3.4 recalculates the binding contribution (E_{score}) with a linear energy score; the angle brackets again represent Boltzmann averages. If a pose has a PLOP energy (E_{pose}) within 5.26 kcal/mol from the lowest one, the energy score (E_{score}) is zero; otherwise, it is 0.58 times the relative energy. The potential sites that have relative score within 1.46 kcal/mol of a site predicted to have the highest reactivity are considered to be a site of metabolism.

Reactivity model: The sites at which a ligand gets metabolized by a P450 enzyme depends not only on whether the atom in question can approach the heme iron center with the correct geometry, but also on the intrinsic chemical reactivity of the site. Assuming that the intrinsic chemical reactivities of the ligand sites are independent of the presence of the enzyme, we estimated the intrinsic reactivities from activation energies of a library of model systems using QM. Since DFT with the B3LYP functional and the 6-31G* basis set has been shown to give high accuracy for relative energies of transitions states,[133] while still allowing for fast calculations, we employed that level of theory for our intrinsic reactivity model. It has been shown that in general, an accurate linear correlation exists between the QM activation energies of hydrogen abstraction reactions with a methoxy radical and the corresponding hydrogen abstraction barriers with an iron-oxo porphyrin species, generally referred to as Compound I in the P450 literatures.[130] In agreement with previous reports,[134,135] we herein investigated the above-mentioned

correlation including aliphatic hydrogen abstraction barriers as well as aromatic ones. As shown in Figure 3.7, we find a good correlation between the methoxy radical and the Compound I based activation barriers for both sp^2 and sp^3 hybridized systems (R^2=0.94), which validates the use of the methoxy radical model to estimate the intrinsic reactivities. Therefore, transition states for methoxy radical based hydrogen abstraction reactions were optimized at the B3LYP/LACVP* level of theory with *Jaguar*[126] for a fragment library consisting of 150 model compounds, 483 distinct hydrogen atoms, and more than 2000 conformations, in order to accurately model all distinct chemical environments. Carbon atom based intrinsic reactivities were then assigned as the Boltzmann weighted activation energies over different transition state conformations. Intrinsic reactivities of the ligand sites were assigned using a simple SMARTS string matching algorithm of the fragment library. Thereby the best matching fragment was determined as the one with (1) the largest number of heavy atoms (2) the most hydrogen atoms and (3) the largest sum of atomic numbers.

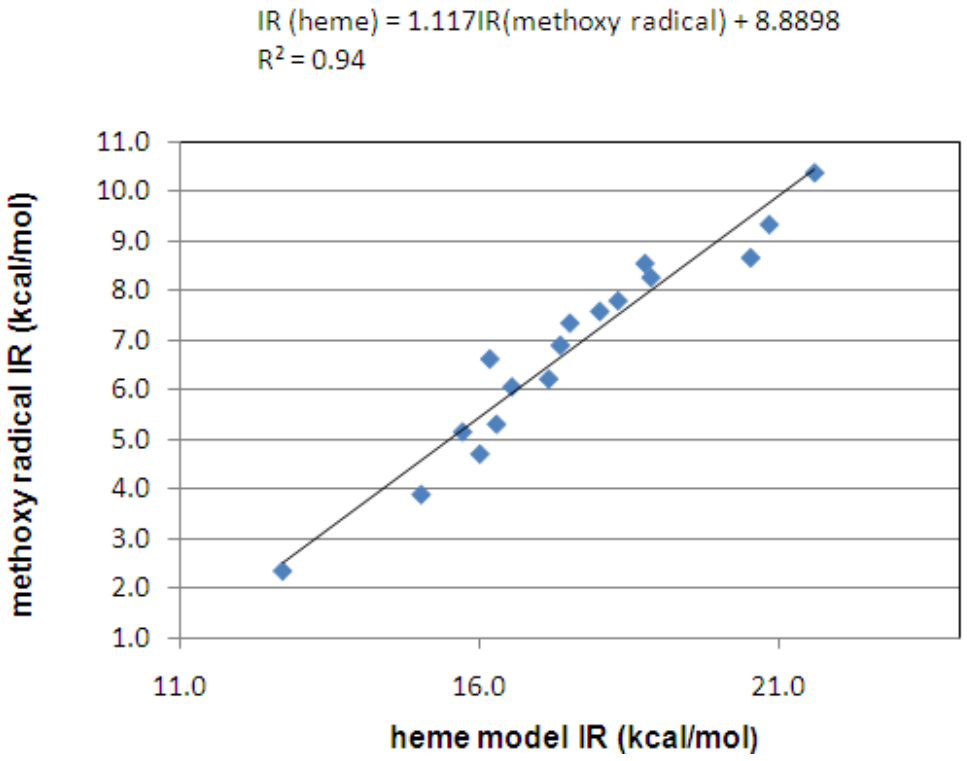


Figure 3.7. Correlation between the intrinsic reactivities calculated with the methoxy radical model and the heme model (17 sites from selected 9 fragment compounds, details are shown in the supporting information).

Preparation of Protein and Ligands

The X-ray crystallographic structure of CYP2D6 was obtained from the Protein Data Bank (PDBID: 2F9Q, 3.0 Å resolution[136]) and contains a well-defined active site above the heme group. We applied the Protein Preparation Wizard (PPW) of Schrödinger Inc. to add hydrogen atoms, optimize the hydroxyl orientation, correct the Gln/Asn/His side-chain orientations, and determine the protonation states of titratable residues. PPW also assigned the bond order of the heme group and the iron oxidation state, which defines the iron atom as Fe^{3+} covalently bonded to the side chain of Cys443. The positions of all hydrogen atoms were optimized with a constraint of 0.3 Å with the OPLS 2005 force field.

A training set of 36 compounds and a test set of 20 compounds were collected from the experimental literature.[37,137] These compounds mainly undergo O-dealkylation and hydroxylation by CYP2D6. The training and test sets contain 774 and 383 heavy atoms, respectively. Details about the data selection are explained in the Supporting Information. All stereoisomers used in the experiments were enumerated as were the protonation states at pH=7.0. All structures were minimized in vacuum using the OPLS 2005 force field, prior to the IDSite calculations.

3.3. Results and Discussions

Tables 3.3 and 3.4 present the summary of our predicted results with the training set and the test set. The data shows that IDSite has high sensitivity and specificity with both IDSite scoring models in predicting the 2D6-mediated metabolism of the 56 compounds; using the Physical IDSite scoring, we achieve high sensitivity (0.83) and high specificity (0.98); using the Fitted IDSite scoring, we can achieve even higher sensitivity (0.94) and similarly high specificity (0.99). With the Fitted IDSite scoring, the results for the training set (sensitivity 0.91 and

specificity 0.99) and test set (sensitivity 1.0 and specificity 0.98) are very similar, indicating that for the fitted model, no overfitting to the training set can be detected.

It is interesting to note that the principal effect of the parameter fitting is to reduce the number of false negatives; the reduction is of similar magnitude in both the training and test sets (there is also some reduction of false positives in the training set, but this is a less prominent result). The principal effect of the parameterization is to take into account the fact that there is some noise in the induced fit calculation energetics, reflected in the 5.26 kcal/mole energy window and scaling factor of 0.58. The noise is a combination of imperfect sampling and residual errors in the continuum solvent free energy model; the parameters suggest that there is a slight overestimation of the relative energetics of poses close in energy. Buffering and scaling the contribution from this term enables a (small) number of secondary sites to be recognized by the model as contributing to the reactivity, without increasing the number of false positives. As noted above, the intrinsic reactivity appears to have less noise associated with it, which is not surprising in view of the fact that it poses a much less demanding sampling challenge.

Table 3.3. Summary of results for the training set.

Symbol	Compound Name	Physical IDSite			Fitted IDSite		
		TP	FP	FN	TP	FP	FN
1	4-methoxyamphetamine	1	0	0	1	0	0
2	Amitriptyline	2	2	0	2	0	0
3	Aprindine	4	0	1	5	0	0
4	Brofaromine	1	0	0	1	0	0
5	Bufuralol	0	1	1	1	0	0
6	Carvedilol	1	0	2	2	0	1
7	Cinnarizine	0	2	1	0	2	1
8	Clomipramine	1	0	1	1	0	1
9	Codeine	1	0	0	1	0	0
10	Desipramine	2	0	0	2	0	0
11	Dextromethorphan	1	0	0	1	0	0
12	Dihydrocodeine	1	1	0	1	0	0
13	Ethylmorphine	1	0	0	1	0	0
14	Flunarizine	1	0	0	1	0	0
15	Fluperlapine	1	0	0	1	0	0
16	Hydrocodone	1	0	0	1	0	0
17	Imipramine	2	0	0	2	0	0
18	Indoramine	1	0	0	1	0	0
19	MDMA	1	0	0	1	0	0
20	Methamphetamine	1	0	0	1	2	0
21	Methoxyphenamine	2	0	0	2	0	0
22	Metoprolol	1	0	1	2	0	0
23	Mexiletine	2	0	1	2	0	1
24	Mianserin	1	0	0	1	0	0
25	Mirtazapine	0	1	1	1	1	0
26	Nortriptyline	1	1	0	1	0	0
27	Ondansetron	2	0	0	1	0	1
28	Paroxetine	1	0	0	1	0	0
29	Perhexiline	2	0	0	2	0	0
30	Propafenone	1	1	0	1	1	0
31	Propranolol	2	2	0	2	1	0
32	Tamoxifen	1	0	0	1	0	0
33	Terfenadine	3	0	0	3	0	0
34	Tiracizine	1	2	0	1	1	0
35	Tropisetron	2	0	1	3	0	0
36	Venlafaxine	1	0	0	1	0	0
	TOTAL	47	13	10	52	8	5

Table 3.4. Summary of results for the test set.

Symbol	Compound Name	Physical IDSite			Fitted IDSite		
		TP	FP	FN	TP	FP	FN
37	Atomoxetine	0	1	1	1	2	0
38	Bicifadine	1	2	0	1	0	0
39	Bupranolol	1	0	0	1	0	0
40	Carteolol	1	1	0	1	0	0
41	Chlorpromazine	1	0	0	1	0	0
42	EMAMC	1	0	0	1	0	0
43	Encainide	1	1	0	1	1	0
44	Harmaline	1	0	0	1	0	0
45	Harmine	1	1	0	1	1	0
46	Ibogaine	1	0	0	1	0	0
47	MAMC	1	0	0	1	0	0
48	MMAMC	1	0	0	1	0	0
49	MOPPP	1	0	0	1	0	0
50	Oxycodone	1	0	0	1	0	0
51	Spirosulfonamide	2	0	0	2	0	0
52	Timolol	2	0	2	4	0	0
53	Tolterodine	0	1	1	1	1	0
54	Tramadol	1	1	0	1	1	0
55	Tyramine	2	0	0	2	0	0
56	Zotepine	1	0	0	1	0	0
	TOTAL	21	8	4	25	6	0

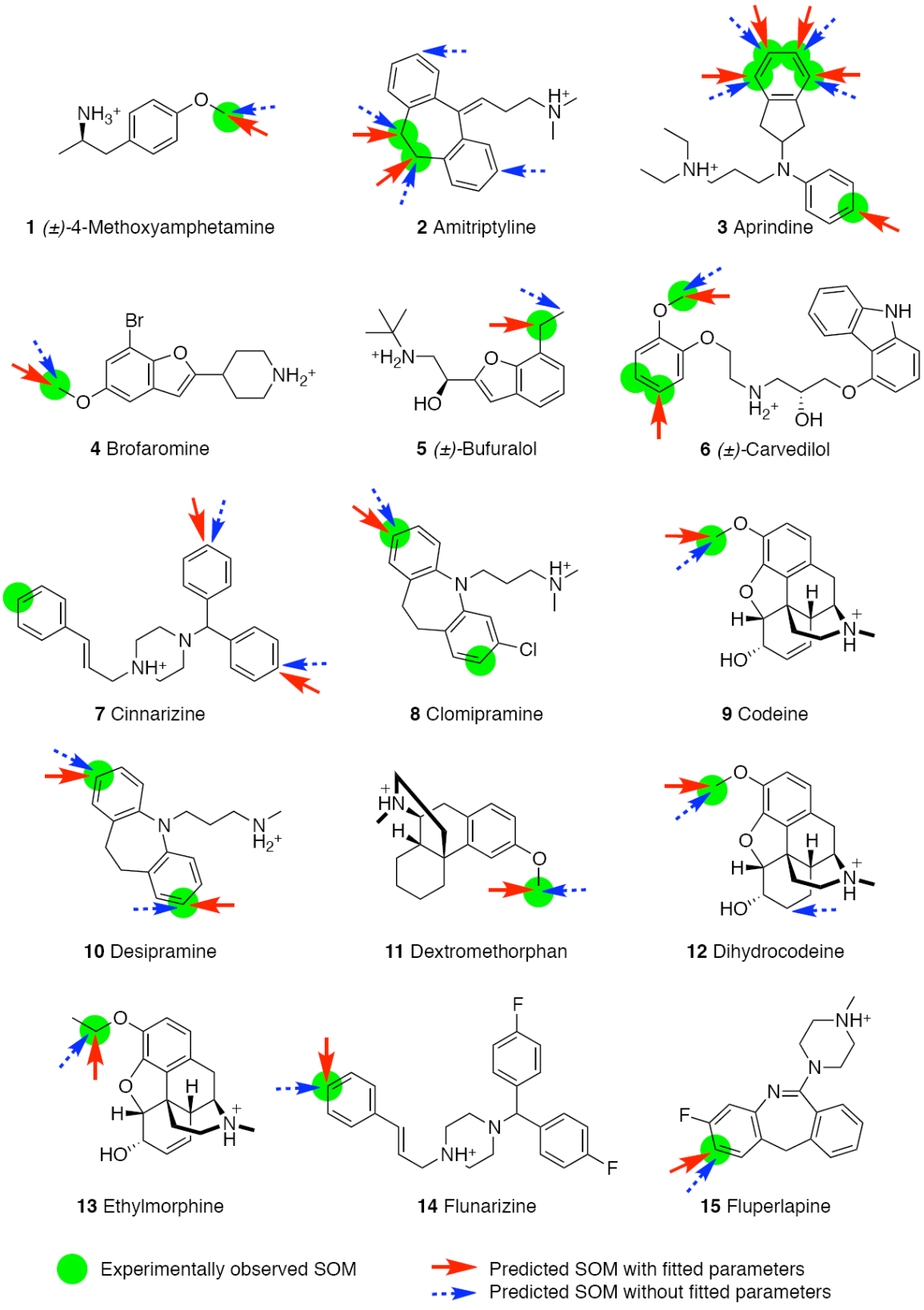


Figure 3.8. IDSite predicted results for the training set.

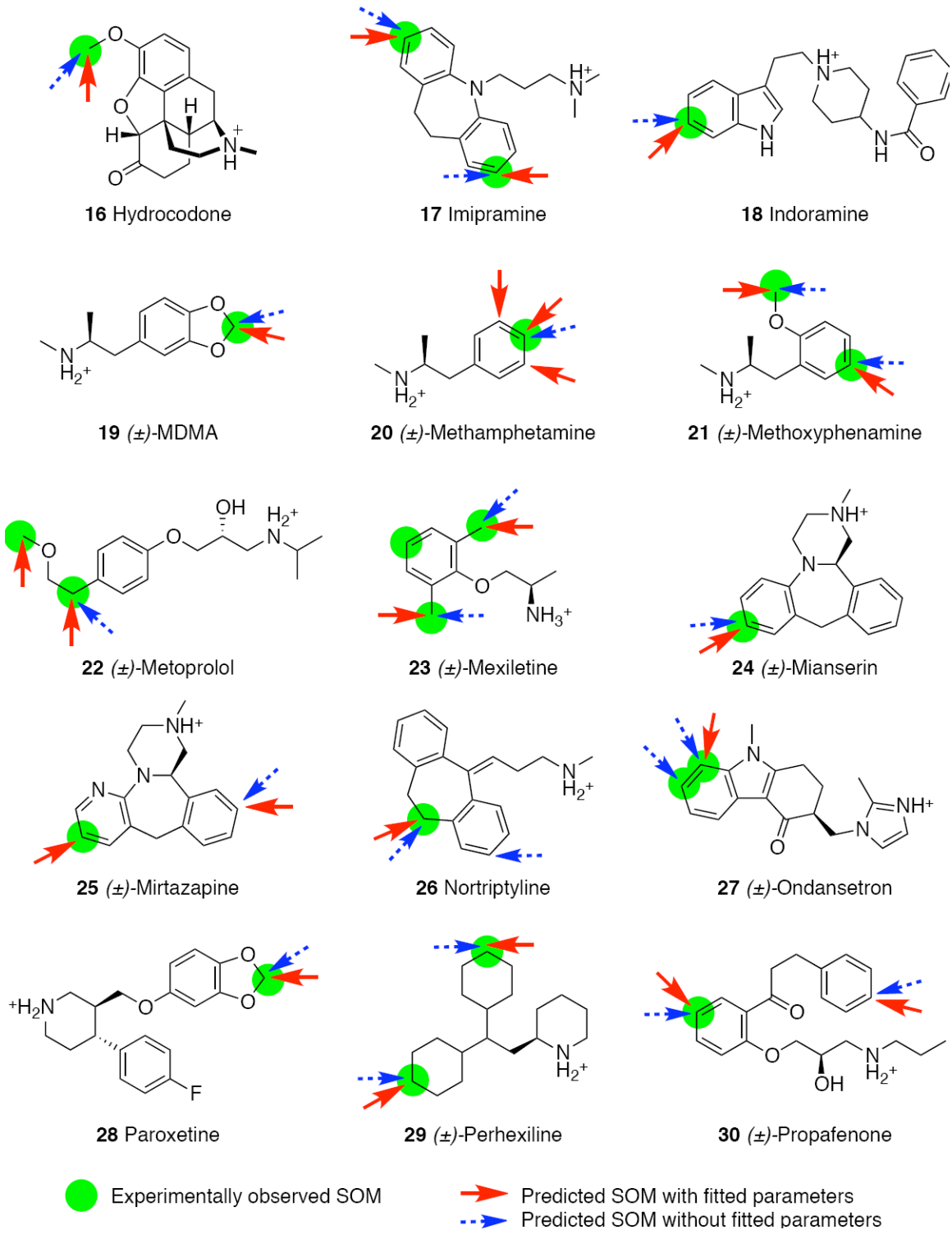


Figure 3.8 (continued). IDSite predicted results for the training set.

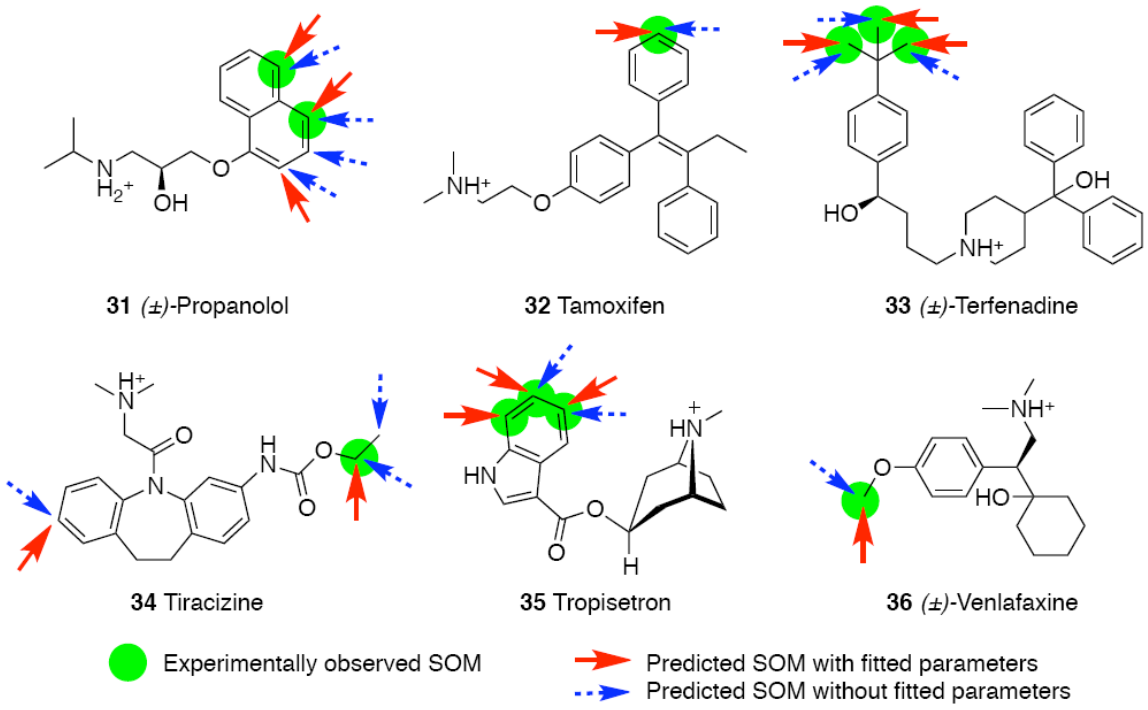


Figure 3.8 (continued). IDSite predicted results for the training set.

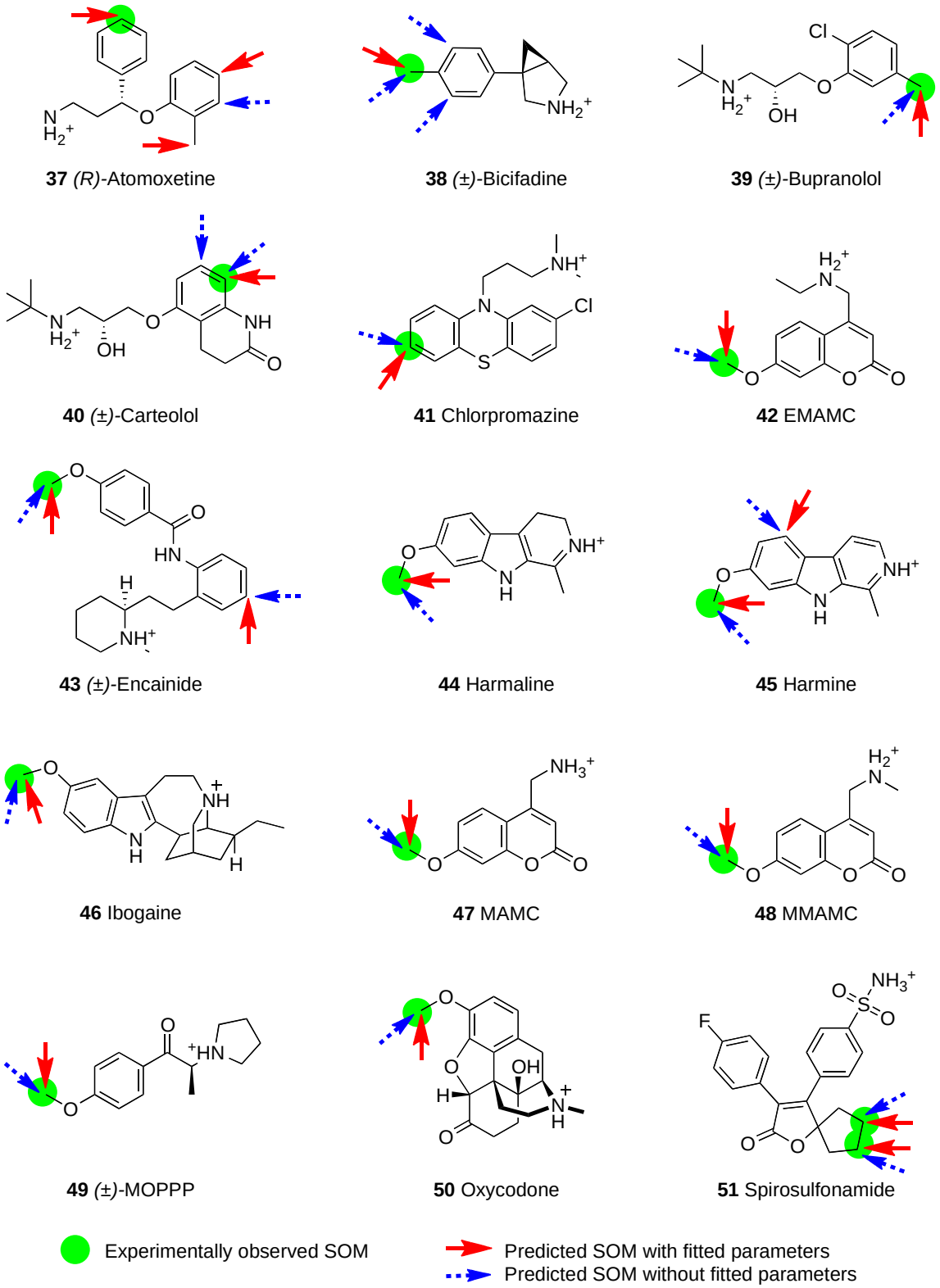


Figure 3.9. IDSite predicted results for the test set.

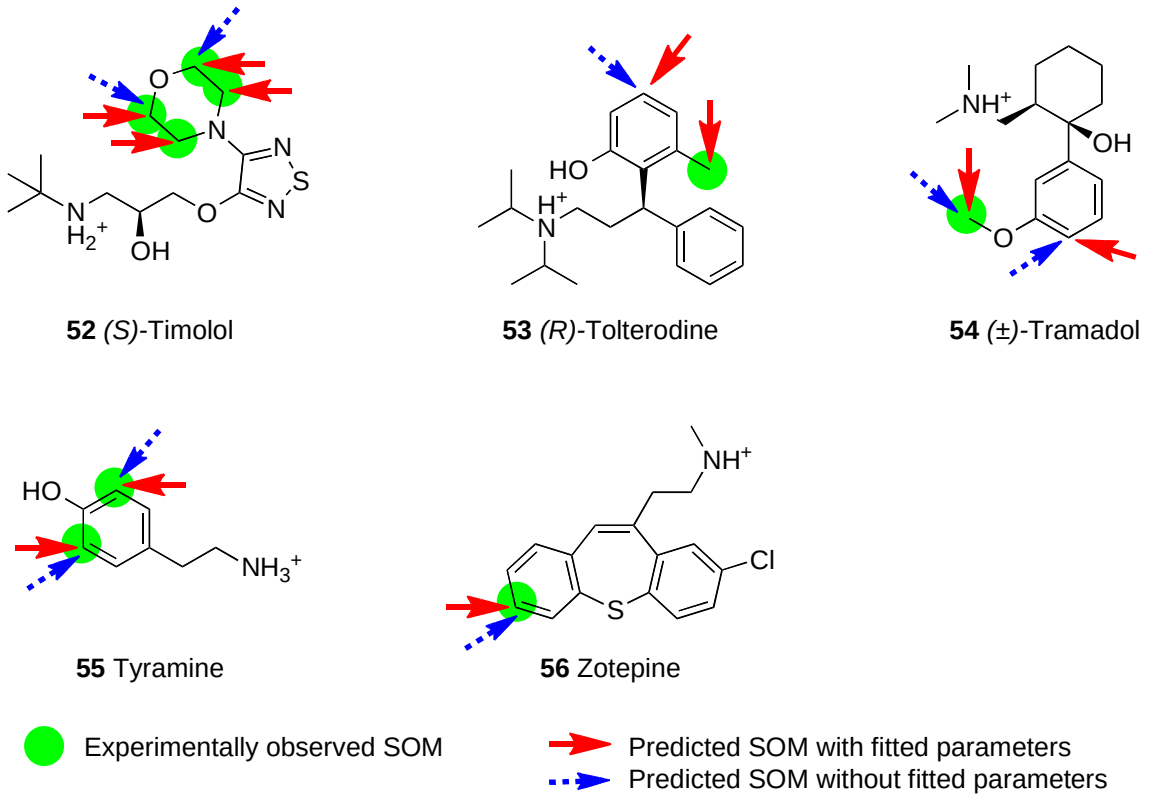


Figure 3.9 (continued). IDSite predicted results for the test set.

A second question of interest is whether the various intensive sampling components of the algorithm actually improve the predictive capability. In order to analyze the importance of each sampling stage in IDSite, ROC (Receiver Operating Characteristic) curves were calculated (Figure 3.10A) to compare three reduced methods using the fitted score to the full method using *physical* and *fitted* scores. As mentioned in the Methods section, each refinement stage performs a constrained minimization, followed by sampling with MCM simulations. After Glide docking, the prediction can be made after minimization in the first refinement stage (referred to as "docking+minimization"), after the sampling in the first refinement stage (referred to as "no Ref2"), or after the minimization in the second refinement stage (referred to as "no sampling in Ref2"). Higher energy cutoff (150 kcal/mol, instead of 24 kcal/mol) and distance cutoff (8.0 Å, instead of 2.6 Å for sp^3 and 2.08 Å for sp^2 hybridized atoms) are adjusted for the methods of "docking+minimization" and "no Ref2". To draw the ROC curves, the scoring cutoff

(4.75 and 1.46 kcal/mol are used for the results shown in Table 3.3 and 3.4 for the *physical* and *fitted* scores, respectively) is varied at 0.5 kcal/mol interval from 0.0 to 100 kcal/mol, which represent the true positive rate (y-axis) and the corresponding false positive rate (x-axis) of the methods. True positive rate and false positive rate are calculated according to Eq. 3.5,

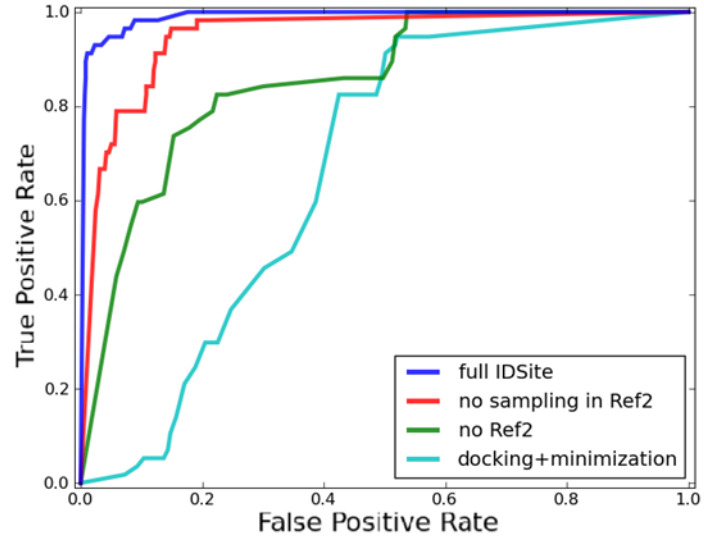
$$True\ positive\ rate = \frac{number\ of\ true\ positives}{number\ of\ SOMs\ observed\ in\ experiments}$$

$$False\ positive\ rate = \frac{number\ of\ false\ positives}{number\ of\ nonSOMs\ observed\ in\ experiments}$$

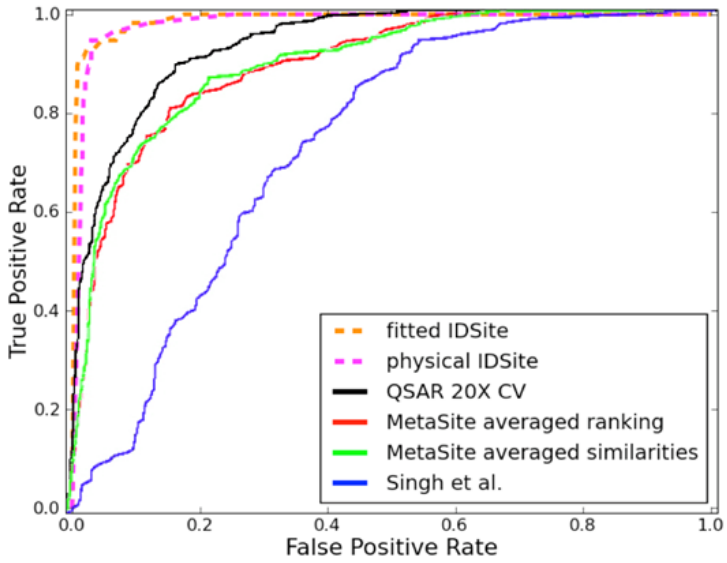
(Eq. 3.5)

where true positives are the SOMs (sp^2 and sp^3 carbon atoms which undergo hydroxylation or O-dealkylation) identified by experiments as well as predicted correctly by IDSite, and the false positives are non-SOMs (nonhydrogen atoms) but mispredicted by IDSite as hydroxylated/dealkylated by CYP2D6. As currently we mainly focus on the typical CYP2D6-mediated hydroxylation and O-dealkylation involving sp^2 and sp^3 carbon atoms with bonded hydrogen atoms, those sites (carbon atoms or heteoatoms) which potentially undergo other metabolic reactions such as N-dealkylation and oxidation are currently considered as non-SOMs in our preliminary study.

The ROC curves in Figure 3.10A indicates that at the same false positive rate (sensitivity), the false positive rate decreases with more sampling and the full IDSite method always has the lowest false positive rate (the highest specificity) with both scoring models. It is interesting that the *physical* score derived from the basic physical chemistry model is very close to the *fitted* score. For the reduced methods, there is an obvious trend that increasing the sampling efforts yields substantially higher specificity at each stage. This means that using the IDSite scoring models in conjunction with binding requirements, sufficient sampling in IDSite can specifically identify the sites metabolism observed by the experiments.



(A)



(B)

Figure 3.10. (A) ROC curves comparing the full IDSite method to the reduced methods. **(B)** ROC curves superimposed on the results of Sheridan *et al.*[41]

In Figure 3.10B we compare the Physical IDSite and Fitted IDSite results to results from Sheridan *et al.*,[41] who evaluated true positive and false positive rates, using the same ROC metric that we employ, for their test set of CYP2D6 ligands. The test set employed in ref. [41] is different in detail from the one we use here, but the types of ligands in both test sets are similar based on examples of test set

molecules given in ref.[41]. Hence, while the comparison is not completely rigorous, it is a reasonable way to estimate relative performance. It can be seen that given the caveat above, both Physical IDSite and Fitted IDSite substantially outperform both MetaSite and the in house Merck QSAR-based approached plotted in Figure 3.10B. To recover 90% of true positives, the QSAR method included roughly 20% of false positives, whereas MetaSite included 40% of false positives. In contrast, IDSite incorporated only ~1% of false positives. This is a qualitative transformation of performance that has significant implications for use in drug discovery applications, as does the availability of a predicted three dimensional structure that is likely to be quite accurate.

So far, only the apo enzyme structure of CYP2D6 has been determined by X-ray crystallography. In order to investigate the capability of IDSite in modeling the induced-fit effects and understand the effects of the hierarchical sampling, several compounds of various sizes and flexibility were selected to analyze the structural and energetic changes at each stage.

It is very common that the poses from docking that have the SOM close to the ferryl oxygen are not among the top poses considered by *Glide* SP scoring. For example, the pose with the shortest distance (1.8 Å) is ranked 6th in the case of 4-methoxyaphetamine; the pose (1.4 Å) that leads to prediction of O-demethylation is ranked 20th for the case of metoprolol. Further, it is also possible for some cases (e.g. fluperlapine) that none of the poses have the SOM close enough to the ferryl oxygen. Therefore, it is very difficult to make specific predictions with only a small distance cutoff and a few top poses from docking. In order to improve the sensitivity as well as the specificity of the predictions, it appears to be necessary to employ the refinement stages.

Focusing on the distance between the site(s) of metabolism observed experimentally, we investigated the Boltzmann averaged energy and distance from the site(s) to the ferryl oxygen over all the poses sampled at any even numbered step. Given the strong harmonic constraints applied in the refinement stages, the distance change is generally relatively small as expected. The energy change in the

first refinement is usually small ranging from 4 to 25 kcal/mol. However, the energy change during the second refinement stage is quite different for small ligands as compared to large ligands. For ligands as small as 4-methoxyaphetamine, the energy of the poses fluctuated within the range of 12 kcal/mol and the lowest energy structure was obtained at the early steps. In contrast, the energy can decrease by more than 60 kcal/mol during the sampling of the second refinement stage for flexible or bulky ligands such as fluperlapine. For such cases it is often not until the end of the simulation that the low energy structure is sampled. This implies that the second refinement plays an important role in optimizing the structure for bulky or flexible compounds.

Skipping the second refinement, about 40% of the compounds (24/56) in the training set have the same results as obtained from the full protocol and most of them are small compounds like 4-methoxyaphetamine, MDMA, MAMC, etc. This observation is consistent with our discussion above that links the need for extended refinement to the presence of large, bulky ligands where protein induced-fit effects are significant, and where optimization of the free energy of the reactive binding complex can pose great difficulties due to various types of energy barriers and additional degrees of freedom to explore in the ligand.

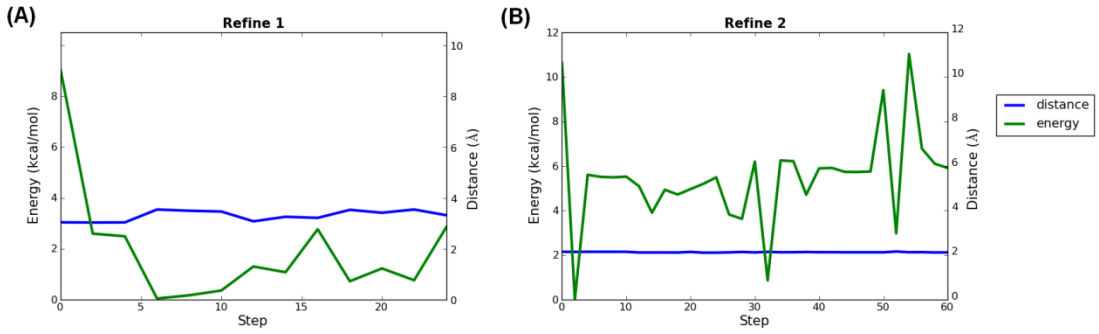


Figure 3.11. The energy and distance (constrained atom to the ferryl oxygen) changes during the MCM simulation during the first **(A)** and the second **(B)** refinements for 4-methoxyaphetamine.

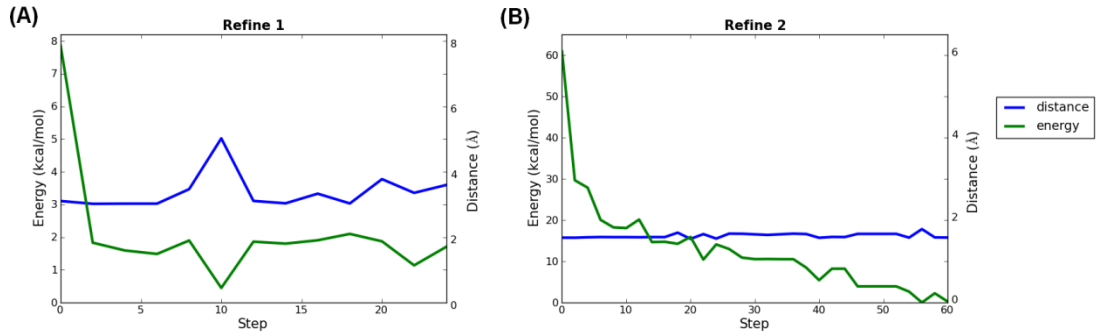


Figure 3.12. The energy and distance (constrained atom to the ferryl oxygen) changes during the MCM simulation during the first **(A)** and the second **(B)** refinements for dextromethorphan.

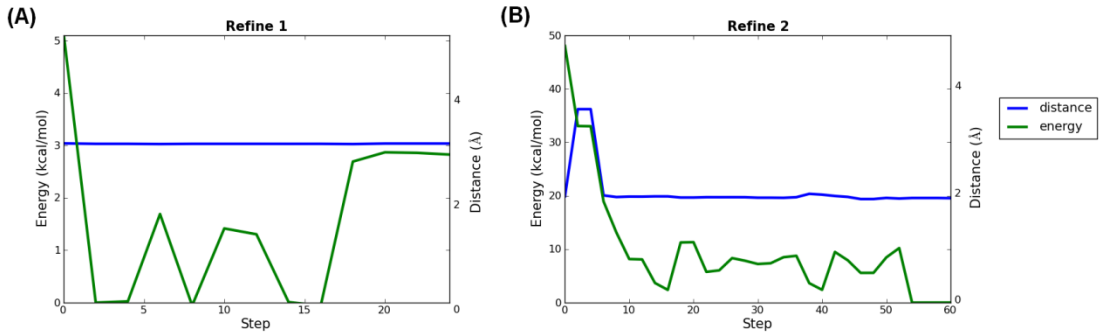


Figure 3.13. The energy and distance (constrained atom to the ferryl oxygen) changes during the MCM simulation during the first **(A)** and the second **(B)** refinements for fluperlapine.

Analysis of Induced-fit Effects

P450 enzymes are believed to have high flexibility in adjusting their active site to accommodate a large variety of substrates. In order to model such induced fit effects, sufficient sampling provided by the two refinement stages of IDSite is critical as demonstrated in the previous section. In order to further investigate the capability of IDSite in modeling induced-fit effects, we calculated the average absolute change for each dihedral angle of the protein side chains in the binding box in comparison to the minimized crystal structure of CYP2D6. The largest change of all the chi angles for each residue is used to represent the change for that residue. Figure 3.14 illustrates the induced-fit effects by showing the largest change for each residue. 10 of the 18 residues in the binding box have changes greater than 30 °. This shows that

IDSite is able to model induced-fit effects required to correctly identify the "bio-active" conformation of the ligands by changing the side-chain orientations in the active site. Phe120 and Phe483 with bulky side chains have changes as large as 40 °and 60 °, respectively. However, the magnitude of their induced fit effects depends strongly on the ligand size. Between these two Phe residues in the binding box, Met374 has the most significant change (108 °) because a small rotation in the Phe side chains can cause a big adjustment in Met374. Compared to the large change of Glu216 (88 °), the change of Aps301 (38 °) is relatively smaller due to the shorter side chain.

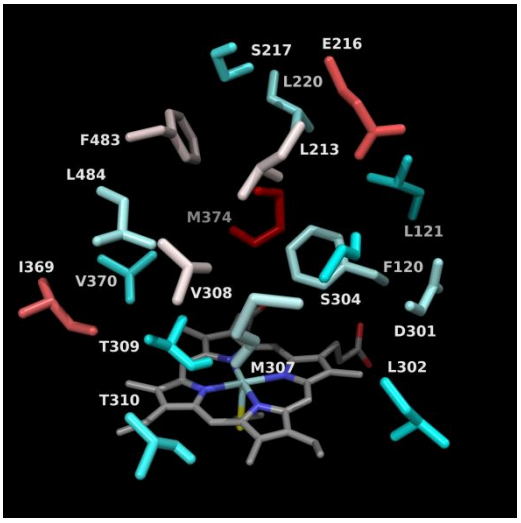


Figure 3.14. Illustration of the induced-fit effects modeled by IDSite. Cyan-white-red scheme is used to show the side chains from the least changed to the most changed, defined as the maximum mean absolute dihedral angle change for each residue.

The above mentioned trends are illustrated in Figures 3.15-3.17, which compare the docked structures leading to the SOM of 4-methoxyamphetamine (PMA), fluperlapine, and metoprolol to the crystal structure of the apo-enzyme minimized with the VSGB 2.0 energy

model. Analogous figures can be found for all our predictions in the supplementary information. One striking example of induced-fit effects involves Phe120. For small ligands such as PMA the benzene ring conformation of Phe120 changes only slightly (Figure 3.15) while it has to move out of the way for larger ligands such as fluperlapine (Figure 3.16) or metoprolol (Figure 3.17), therefore rotating by almost 90 °. Interestingly, for compounds with multiple sites of metabolism, such as metoprolol (Figure 3.17), different binding modes leading to different SOMs have very different conformations of the Phe120 side chain as well. The structures produced by IDSite clearly highlight the importance of induced fit effects for CYP2D6 metabolism and provide an explanation for why it is difficult to accurately predict SOMs with a rigid receptor model.

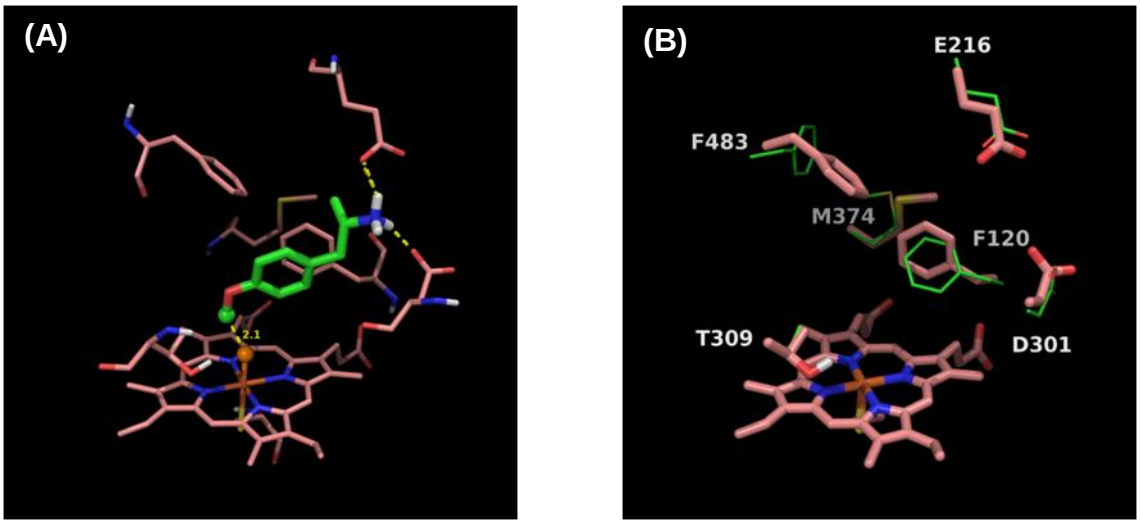


Figure 3.15. (A) The lowest energy pose in the second refinement stage for 4-methoxyaphetamine. Orange sphere = "dummy" ferryl oxygen, green sphere = experimental and predicted SOM. **(B)** Comparison of side chains important for induced-fit effects. Crystal structure (green, PDBID: 2F9Q) minimized with the VSGB 2.0 model and superimposed onto the lowest energy pose with 4 methoxyamphetamine (salmon). Large dihedral changes are seen for Asp301 (Δchi2, 121°), Met374 (Δchi3, 114°), and Phe483 (Δchi1, 60°).

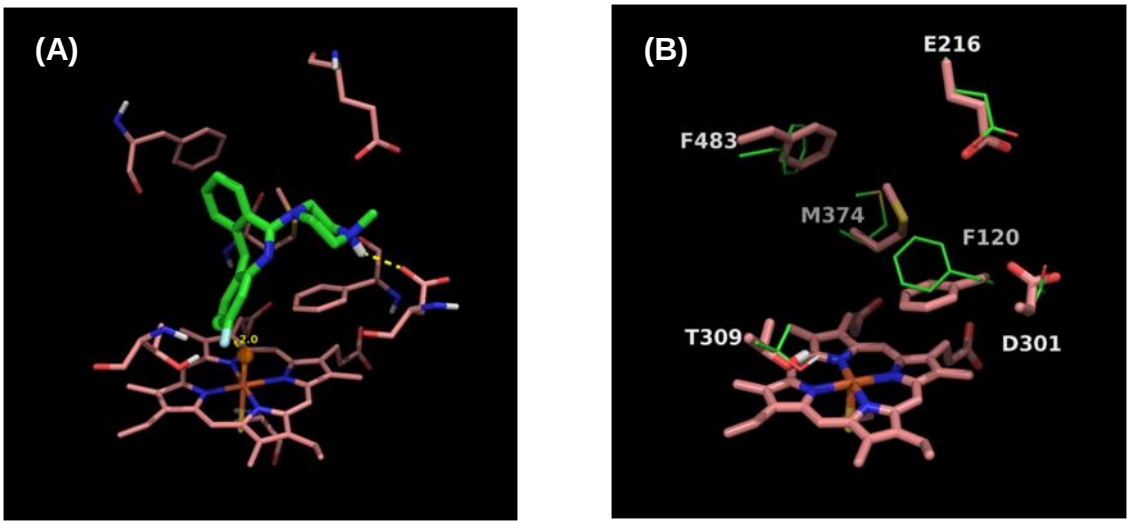


Figure 3.16. (A) The lowest energy pose in the second refinement stage for fluperlapine. Orange sphere = "dummy" ferryl oxygen, green sphere = experimental and predicted SOM. **(B)** Comparison of side chains important for induced fit effects. Crystal structure (green, PDBID: 2F9Q) minimized with the VSGB 2.0 model and superimposed onto the lowest energy pose with Fluperlapine (salmon). Large dihedral changes are seen for Phe120 (Δchi2, 73°), Glu216 (Δchi1, 60°), Asp301 (Δchi2, 64°), Met374 (Δchi3, 105°), and Phe483 (Δchi2, 94°).

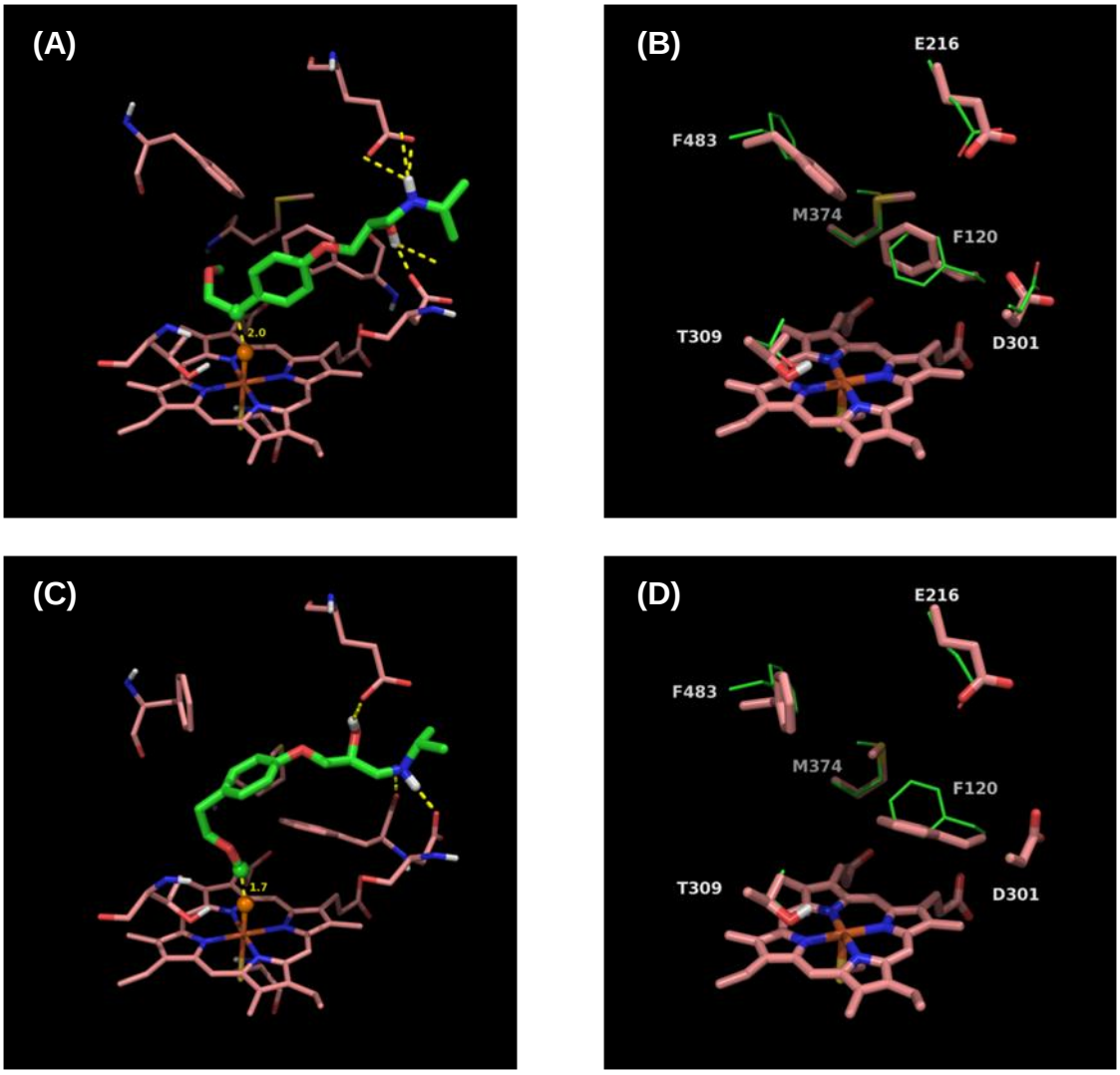


Figure 3.17. (A) The lowest energy poses in the second refinement stage for metoprolol benzylic hydroxylation. **(B)** Comparison of side chains important for induced fit effects for metoprolol benzylic hydroxylation. **(C)** The lowest energy poses in the second refinement stage for metoprolol O-dealkylation. **(D)** Comparison of side chains important for induced fit effects for metoprolol O-dealkylation. For **(A)** and **(C)** orange spheres = "dummy" ferryl oxygen, green spheres = experimental and predicted SOMs. For **(B)** and **(D)** crystal structure (green, PDBID: 2F9Q) minimized with the VSGB 2.0 model and superimposed onto the lowest energy poses with metoprolol (salmon). For benzylic hydroxylation, large dihedral changes are seen for Glu216 (Δchi1, 60°), Asp301 (Δchi2, 66°), Met374 (Δchi3, 112°), and Phe483 (Δchi1, 40°); for O-dealkylation, large dihedral changes are seen for Phe120 (Δchi2, 67°), Glu216 (Δchi2, 50°), and Phe483 (Δchi2, 194°)

Importance of Structural Effects in Determining SOMs

The two main competing factors in determining the SOMs with P450 enzymes are the intrinsic reactivities of the ligand sites and the geometric fit of the ligand in the active site. As mentioned in the Methods section, IDSite considers both of these effects in determining the SOMs, which enables it to select the correct SOM even for difficult cases, where the intrinsic reactivity favors a site that is not experimentally observed to be oxidized. For these cases, the structural fit of the ligand with the receptor (i.e., how easily the ligand site can reach the ferryl oxygen) mainly determines the SOM. Therefore, the structures and energies of the poses, with consideration of the receptor, have to be utilized. Three cases are used here to demonstrate the role of a receptor (CYP2D6) in determining the sites of metabolism.

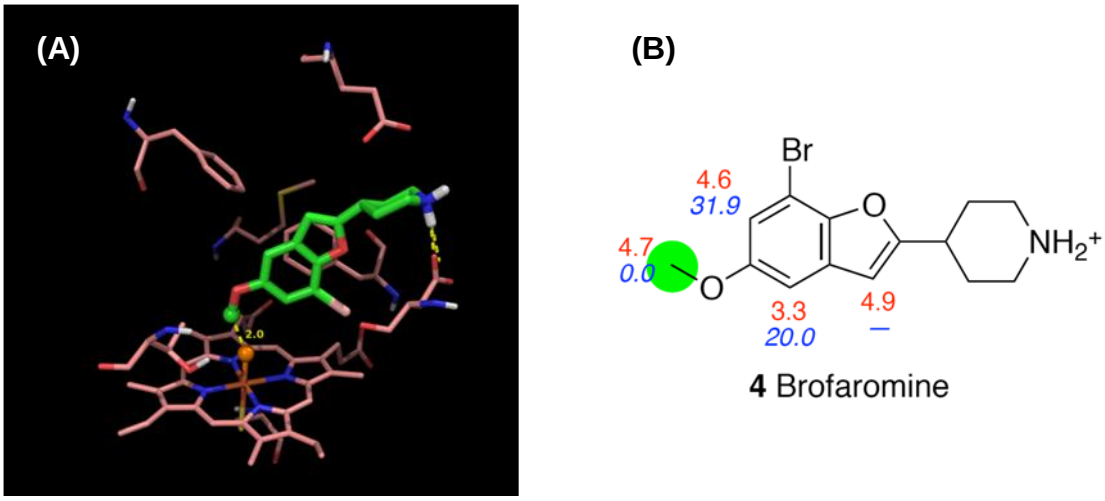


Figure 3.18. (A) The lowest energy pose in the second refinement stage for brofaromine. Orange sphere = "dummy" ferryl oxygen, green sphere = experimental and predicted SOM. **(B)** Intrinsic reactivities (red) for each site and the relative energy (blue) of the poses with the corresponding site constrained to the ferryl oxygen. The SOM observed experimentally is marked with a green circle.

The first case is brofaromine, for which experiments show that the major metabolic pathway is O-demethylation mediated by CYP2D6.[138] The intrinsic reactivity of the site of metabolism (4.7 kcal/mol) is very close to those of sites on the aromatic rings (non sites of metabolism, 3.3-4.9 kcal/mol) (Figure 3.18). Due to the receptor geometry, it is impossible for the atoms on the furan ring to approach the ferryl oxygen while still forming the required salt bridge with either Glu216 or Asp301. Therefore, no qualified poses were found leading to a reaction on the furan ring. Although we found qualified poses for all the sites on the benzene ring, those poses are all strongly disfavored energetically by more than 20 kcal/mol. This indicates that taking the interactions between the ligand and the receptor into account, IDSite is able to make the prediction of the SOM for brofaromine in good agreement with the experimental observation.

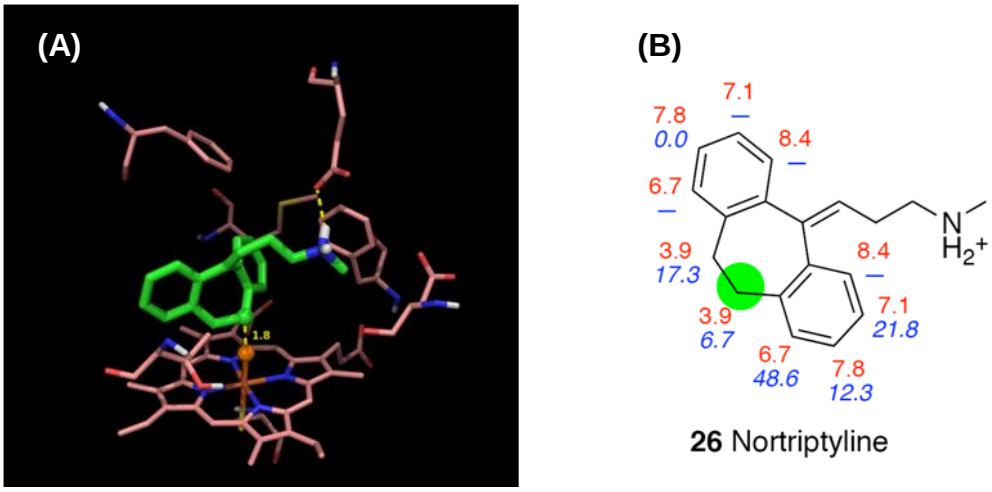


Figure 3.19. (A) The lowest energy pose in the second refinement stage for nortriptyline. Orange sphere = "dummy" ferryl oxygen, green sphere = experimental and predicted SOM. **(B)** Intrinsic reactivities (red) for each site and the relative energy (blue) of the poses with the corresponding site constrained to the ferryl oxygen. The SOM observed experimentally is marked with a green circle.

A second interesting case is nortriptyline, since the two sites on the 7-membered aliphatic ring are difficult to distinguish only with their intrinsic reactivity as they are almost equally reactive. However, experiments show that only the *(E)*-10 site of nortriptyline is metabolized.[139] The poses generated by IDSite with the *(Z)*-10 site close to the ferryl oxygen are all at least 10 kcal/mol higher in energy compared to the poses with the *(E)*-10 atom close to the ferryl oxygen. Such an energy gap is large enough for IDSite to correctly determine the *(E)*-isomer as the only metabolite. While structural effects are therefore clearly very important to determine nortriptylene's SOM, the intrinsic reactivities also play a key role. This is again nicely illustrated with the example of nortriptyline, where a simply structure based method (without considering intrinsic reactivities) would predict the SOM as being an aromatic hydroxylation due to the favorable energy of the corresponding poses. Therefore, IDSite is able to correctly balance the subtle effects stemming from intrinsic reactivity and structural fit.

Methoxyphenamine is another case where the joint effects of intrinsic reactivity and the structural fit lead to the correct predictions. Methoxyphetamine is metabolized through O-demethylation and aromatic hydroxylation mediated by CYP2D6.140 These two sites not only have very close intrinsic reactivities (5.7 and 6.3 kcal/mol, Figure 3.20), but their lowest energy poses also have very similar energies. The non-SOMs are not selected by IDSite because of either disfavorable intrinsic reactivity or high pose energies.

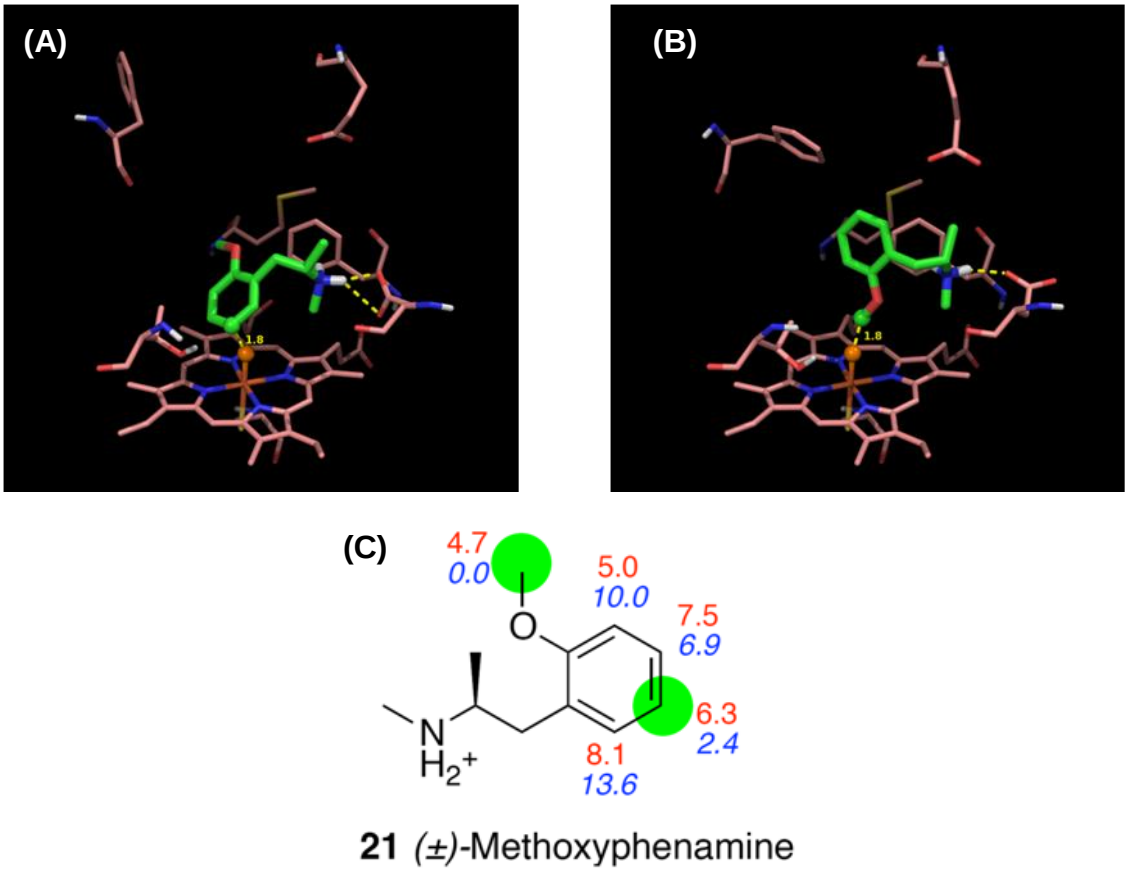


Figure 3.20. The lowest energy pose in the second refinement stage for methoxyphenamine. Orange sphere = "dummy" ferryl oxygen, green sphere = experimental and predicted SOM. **(A)** Aromatic hydroxylation. **(B)** O-demethylation. **(C)** Intrinsic reactivities (red) for each site and the relative energy (blue) of the poses with the corresponding site constrained to the ferryl oxygen. The SOM observed experimentally is marked with a green circle.

Computational Cost

On a single 2.2 GHz AMD Opteron Processor 6174, the average CPU time required for a typical IDSite calculation (e.g. with a compound with 3 rotatable bonds) is about 448 hours, of which about 11% of the time is spent on the first refinement stage and 89% on the second refinement. On 20 such processors the calculation takes 22 hours. The initial Glide docking step on a single processor takes about 10 min. The computational cost of PLOP refinement is proportional to the number of rotatable bonds in the compound.

3.4. Conclusions

A novel approach for the prediction of experimentally observable cytochrome P450 sites of metabolism, IDSite, has been developed and applied it to a data set for the 2D6 P450 isoform. It has shown remarkably high sensitivity and specificity using a structure-based model, representing a major advance as compared to alternatives in the literature, including various types of ligand-based models. The method delivers not only accurate SOM predictions, but also three dimensional structures of the protein-ligand complex, including induced fit effects (which are quite significant), for every SOM identified by the algorithm.

CYP2D6 was selected as the initial target because the binding of a positive nitrogen in the ligand to an acidic group in the protein created an additional constraint that was useful in limiting sampling and achieving reliable poses in the induced fit docking effort. Other important P450 isoforms, such as 2C9 and 3A4, may be more difficult to model in this fashion as they lack such a salt bridge constraint; nevertheless, even if additional sampling effort is required, it should be possible to obtain successful results given the performance of the conformational energy and reactivity models that have seen in the present work. The development of models for additional isoforms, and to additional ligand test sets, is ongoing in our laboratory. Ultimately, predictive use in an active drug discovery project will be required for validation.

3.5. Appendix for Chapter 3

1. Data Set

The details of our data set of CYP2D6 substrates are listed in Table 3.5. Compounds 1-36 were selected from the work of de Groot *et al.*[137] 4 compounds from the data set of de Groot (debrisoquine, GBR-12909, guanoxan, and phenformin) were not included, because the metabolic pathways and the role of CYP2D6 in these pathways were not clear to us.[37,141-143] Compound 37-56 were collected from the review of Wang *et al.*,[37] and further examined by carefully inspecting the original experimental literature. (Table 3.5)

Table 3.5. Details of the data set.

Symbol	Compound name	Major metabolic pathway	Number of SOM	Number of non SOM	Reference
1	(±)-4-methoxyamphetamine	O-dealkylation	1	11	144
2	amitriptyline	benzylic hydroxylation	2	19	145
3	aprindine	aromatic hydroxylation	5	19	146
37	atomoxetine	aromatic hydroxylation	1	18	147
38	(±)-bicifadine	benzylic hydroxylation	1	12	148
4	brofaromine	O-dealkylation	1	17	138
5	(±)-bufuralol	benzylic hydroxylation	1	18	149,150
39	(±)-bupranolol	benzylic hydroxylation	1	17	151
40	(±)-carteolol	aromatic hydroxylation	1	20	152
6	(±)-carvedilol	aromatic hydroxylation O-dealkylation	3	27	153
41	chlorpromazine	aromatic hydroxylation	1	20	154
7	cinnarizine	aromatic hydroxylation	1	27	155
8	clomipramine	aromatic hydroxylation	2	20	156
9	codeine	O-dealkylation	1	21	157
10	desipramine	aromatic hydroxylation	2	18	158
11	dextromethorphan	O-dealkylation	1	19	159
12	dihydrocodeine	O-dealkylation	1	21	160
42	EMAMC	O-dealkylation	1	16	161
43	(±)-encainide	O-dealkylation	1	25	162
13	ethylmorphine	O-dealkylation	1	22	163

14	flunarizine	aromatic hydroxylation	1	29	[155]
15	fluperlapine	7-hydroxylation	1	22	[164]
44	harmaline	O-dealkylation	1	15	[165]
45	harmine	O-dealkylation	1	15	[165]
16	hydrocodone	O-dealkylation	1	21	[166,167]
46	ibogaine	O-dealkylation	1	22	[168]
17	imipramine	aromatic hydroxylation	2	19	[158]
18	indoramine	6-hydroxylation	1	25	[169]
47	MAMC	O-dealkylation	1	14	[161]
19	(±)-MDMA	O-dealkylation	1	13	[170]
20	(±)-methamphetamine	4-hydroxylation	1	10	[171]
21	(±)-methoxyphenamine	5-hydroxylation	2	11	[140]
22	(±)-metoprolol	O-dealkylation benzylic hydroxylation	2	17	[149]
23	(±)-mexiletine	O-dealkylation benzylic hydroxylation aromatic hydroxylation	3	10	[172,173]
24	(±)-mianserin	aromatic hydroxylation	1	19	[174]
25	(±)-mirtazapine	aromatic hydroxylation	1	19	[175]
48	MMAMC	O-dealkylation	1	15	[161]
49	(±)-MOPPP	O-dealkylation	1	16	[176]
26	nortriptyline	benzylic hydroxylation	1	19	[139]
27	(±)-ondansetron	7-hydroxylation 8-hydroxylation	2	20	[177]
50	oxycodone	O-dealkylation	1	22	[178]
28	paroxetine	O-dealkylation	1	23	[179]
29	(±)-perhexiline	aliphatic hydroxylation	2	18	[180]
30	(±)-propafenone	aromatic hydroxylation	1	24	[181]
31	(±)-propranolol	aromatic hydroxylation	2	17	[182]
51	spirosulfonamide	aliphatic hydroxylation	2	25	[128]
32	tamoxifen	aromatic hydroxylation	1	27	[183]
33	(±)-terfenadine	t-Butyl hydroxylation	3	32	[184]
52	(s)-timolol	aliphatic hydroxylation	4	17	[185]
34	tiracizine	O-dealkylation	1	26	[186]
53	(R)-tolterodine	benzylic hydroxylation aromatic hydroxylation	1	23	[187]
54	(±)-tramadol	O-dealkylation	1	18	[188]
35	tropisetron	5-hydroxylation 6-hydroxylation 7-hydroxylation	3	18	[177]

55	tyramine	aromatic hydroxylation	2	8	[189]
36	(±)-venlafaxine	O-dealkylation	1	19	[190,191]
56	zotepine	aromatic hydroxylation	1	20	[192]
	TOTAL		82	1075	

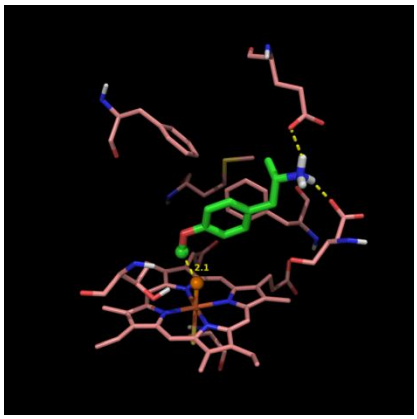
2. Data used to plot Figure 3.7

We show below the model compounds and the data which were used to plot Figure 3.7. Only a doublet spin state was considered for the spin state of Compound I and the transition states.

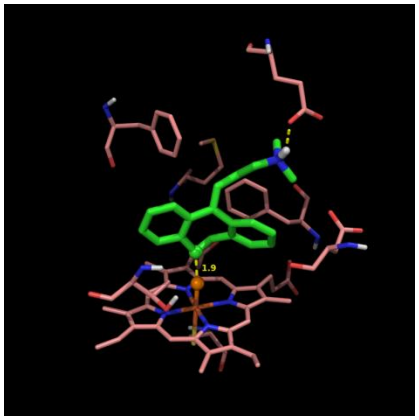
Table 3.6. Comparison of activation energies calculated with the heme model and with the methoxy radical model.

Model compound	Hydroxylated position	Heme model (kcal/mol)	Methoxy radical model (kcal/mol)
benzene		20.51	8.66
toluene	Ortho-	17.15	6.22
	Meta-	18.86	8.27
	Para-	18.00	7.58
	Alpha-	15.72	5.16
anisole	Ortho-	16.29	5.31
	Meta-	18.76	8.55
	Para-	16.01	4.71
	Beta-	16.18	6.63
ethane		21.58	10.37
propane	2-	18.31	7.80
ethanol	1-	12.73	2.36
	2-	17.35	6.90
t-Butylebenzene	Beta-	20.82	9.33
dimethylether		15.03	3.90
dimethylanisole	Meta-	16.54	6.07
	Para-	17.51	7.35

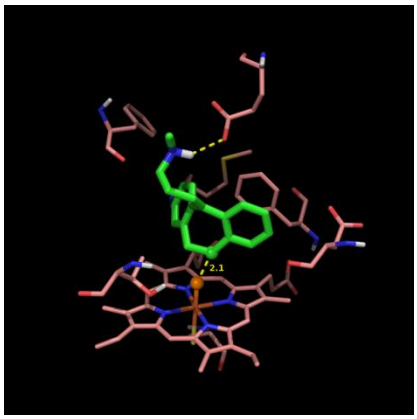
3. Docked Structures used for the correct prediction of SOMs. Only the lowest energy poses that lead to the true positive predictions are shown.



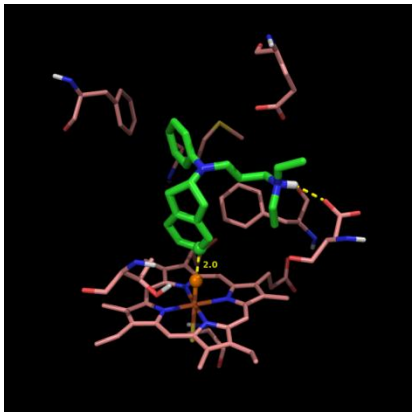
4-methoxyamphetamine



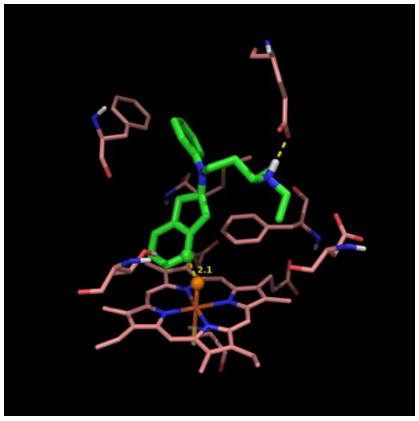
amitriptyline



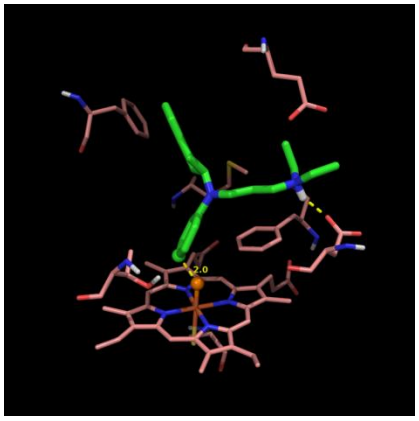
amitriptyline



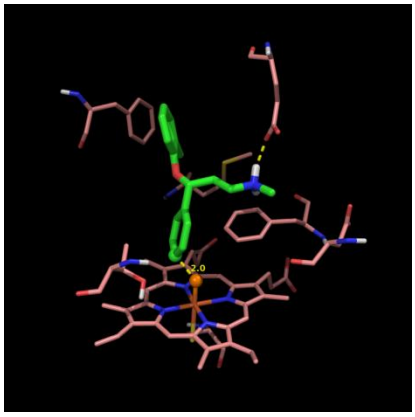
aprindine



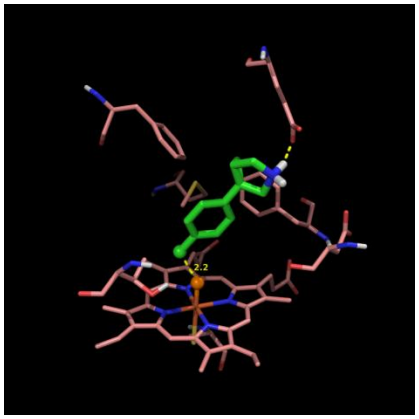
aprindine



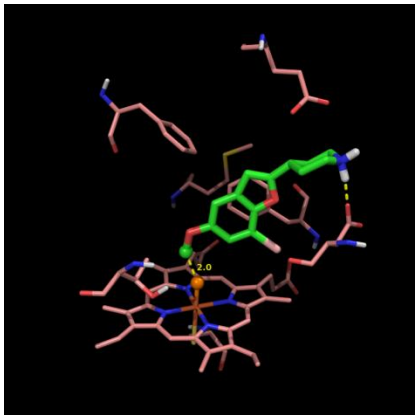
aprindine



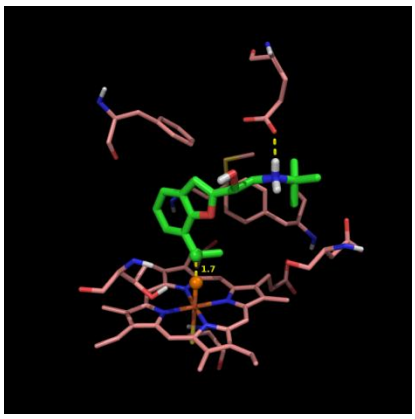
atomoxetine



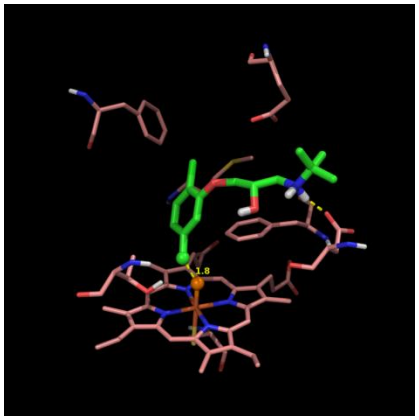
bicifadine



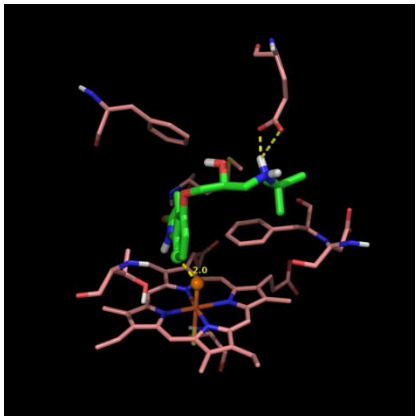
brofaromine



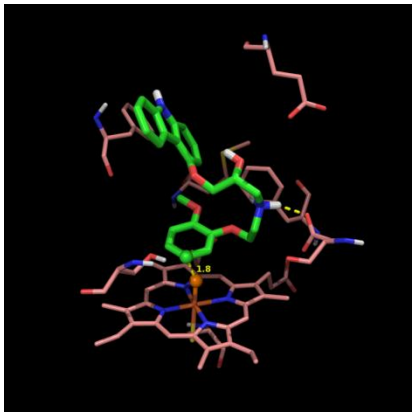
bufuralol



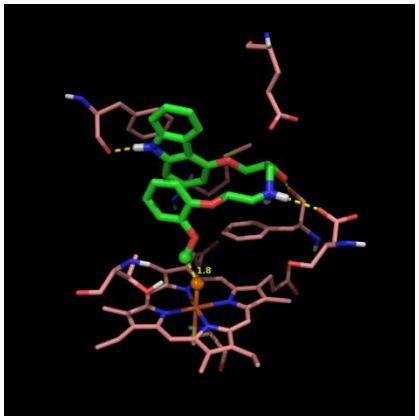
bupranolol



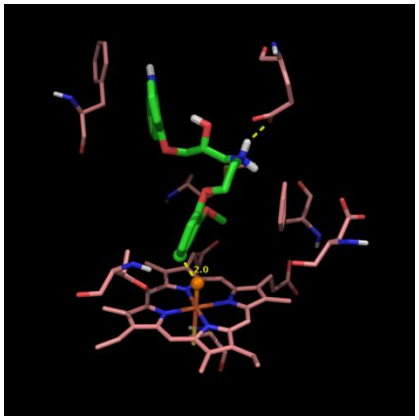
carteolol



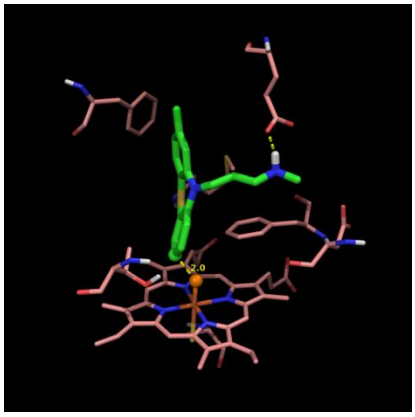
carvedilol



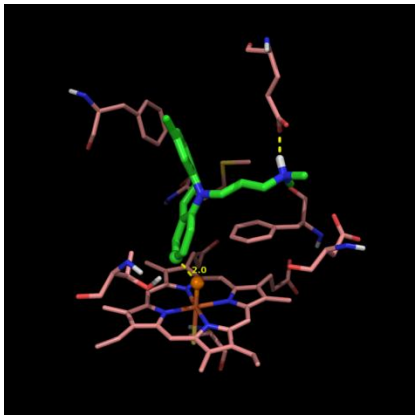
carvedilol



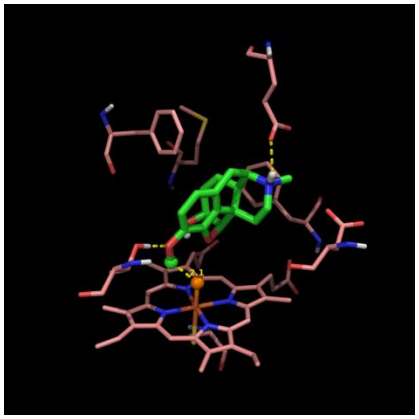
carvedilol



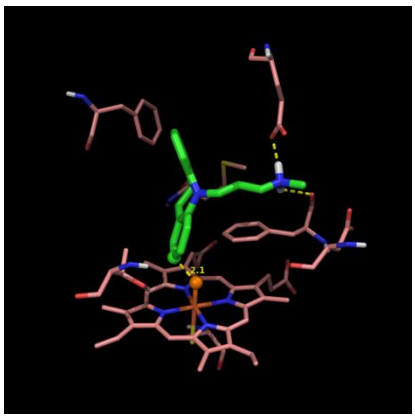
chlorpromazine



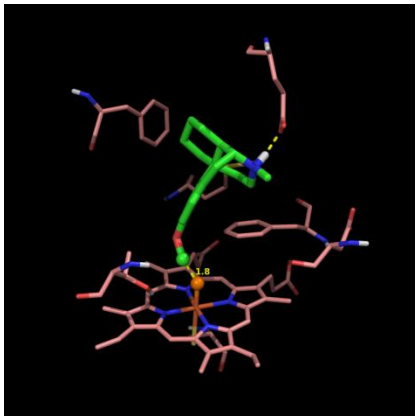
clomipramine



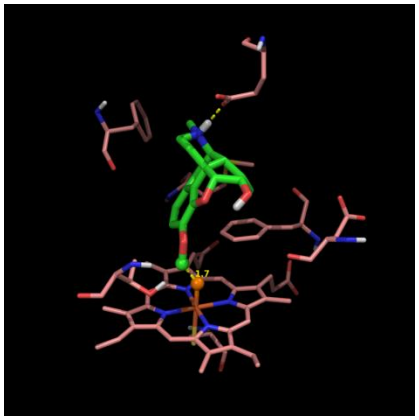
codeine



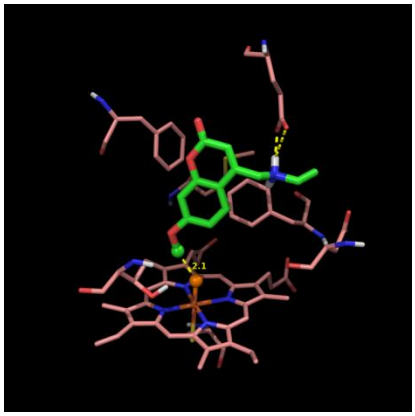
desipramine



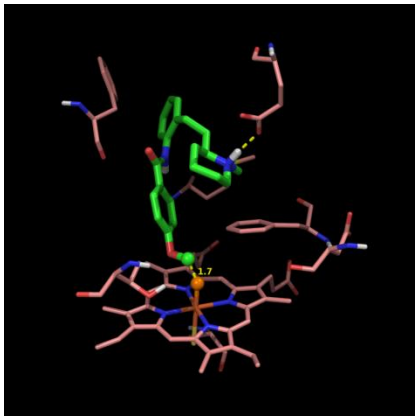
dextromethorphan



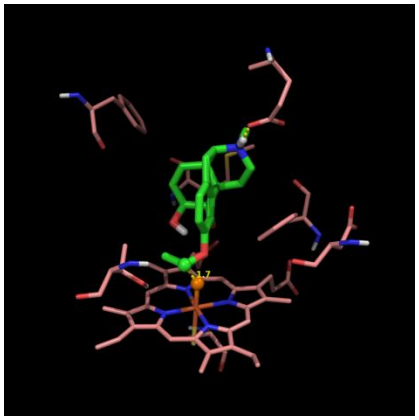
dihydrocodeine



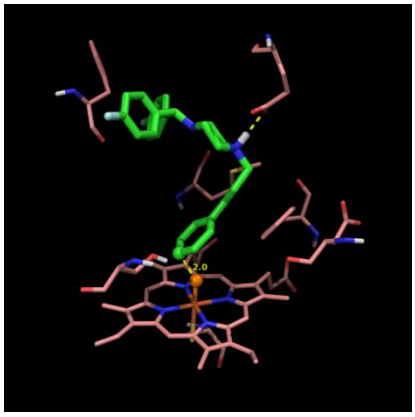
EMAMC



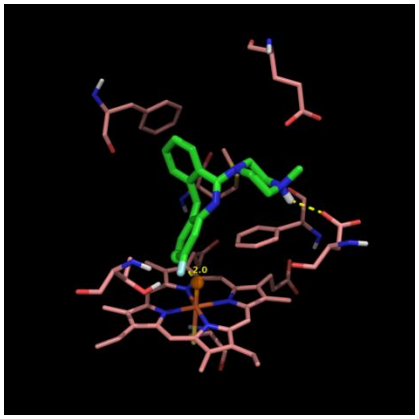
encainide



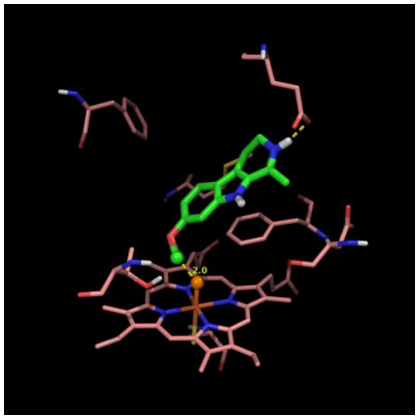
ethylmorphine



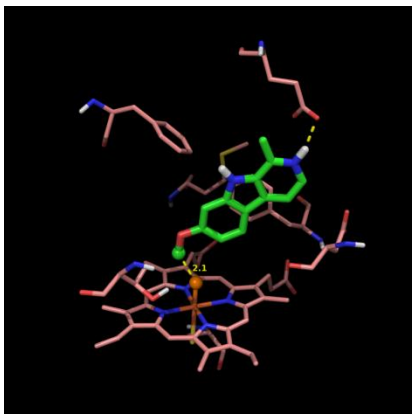
flunnarizine



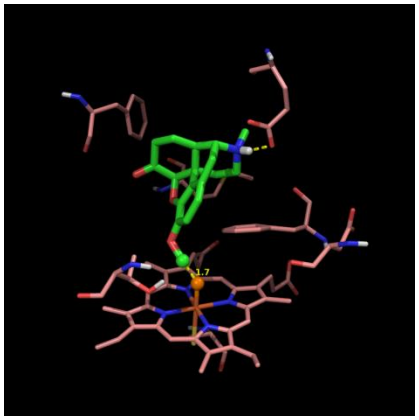
fluperlapine



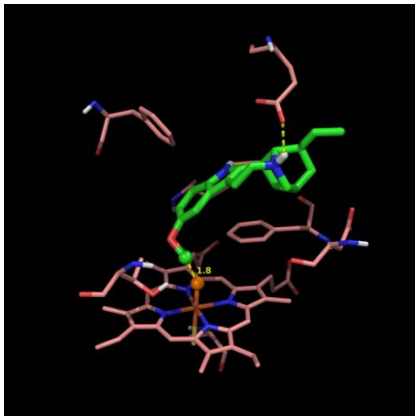
harmaline



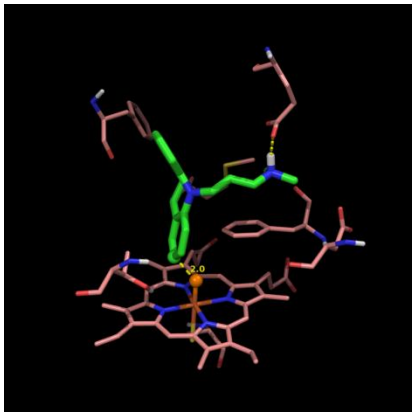
harmine



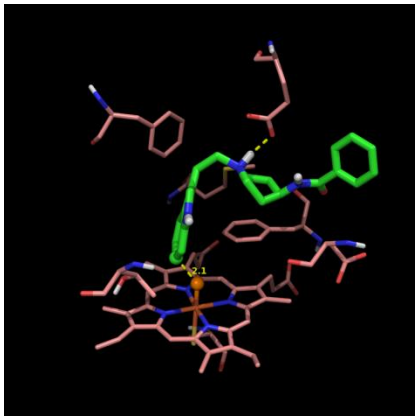
hydrocodone



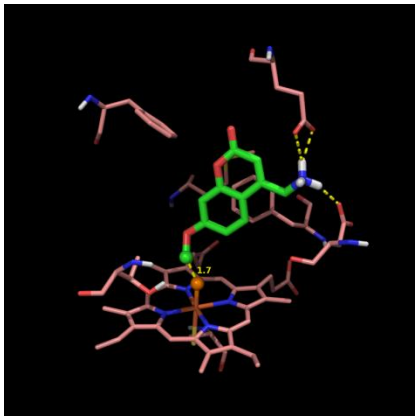
ibogaine



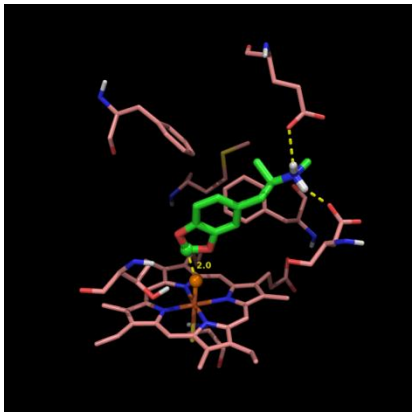
imipramine



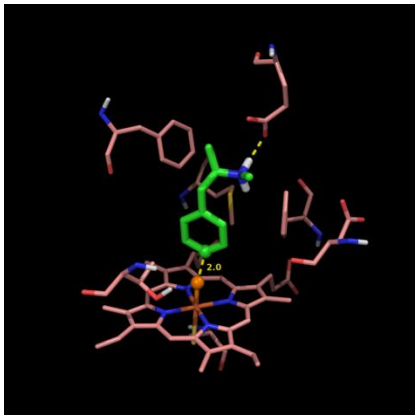
indoramine



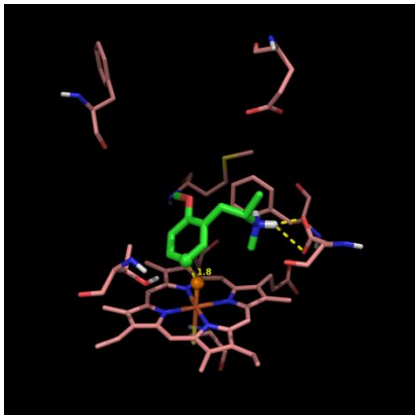
MAMC



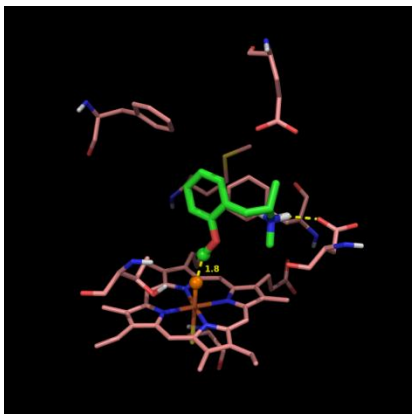
MDMA



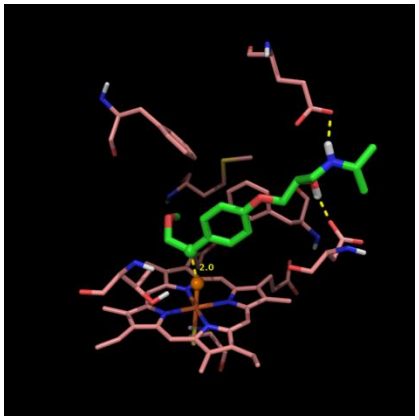
methamphetamine



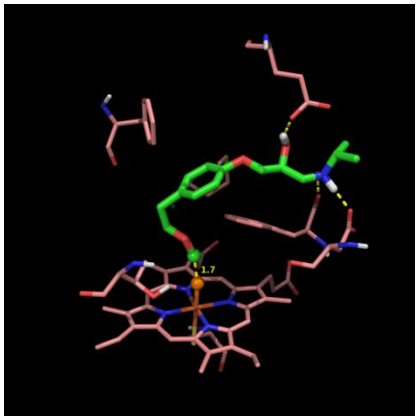
methoxyphenamine



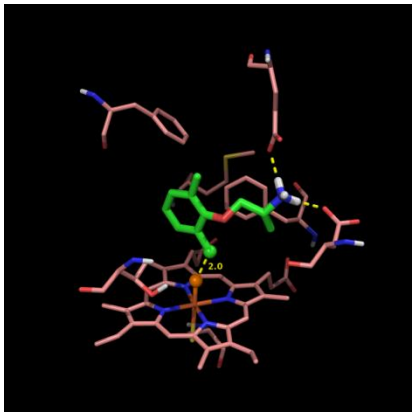
methoxyphenamine



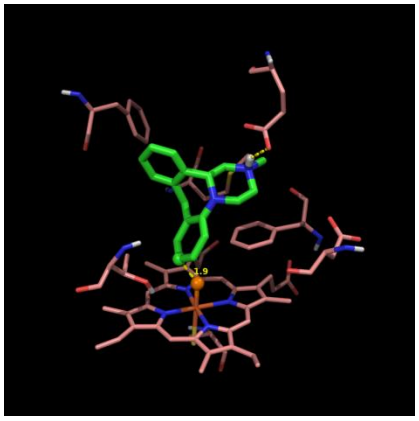
metoprolol



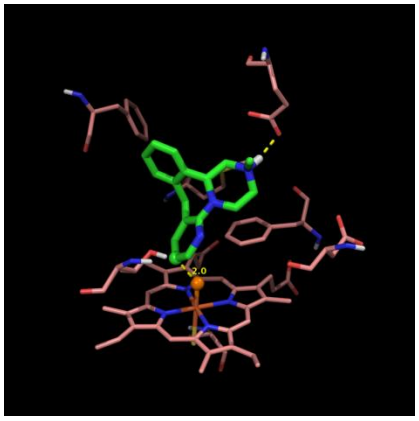
metoprolol



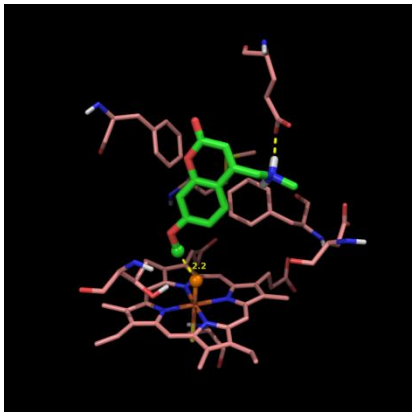
mexiletine



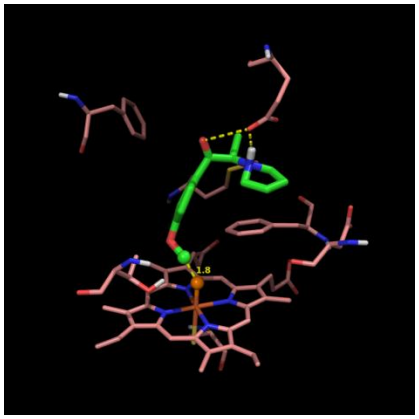
mianserin



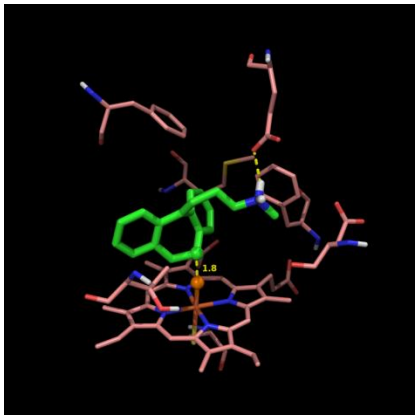
mirtazapine



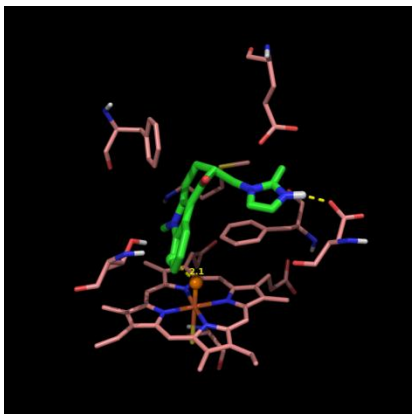
MMAMC



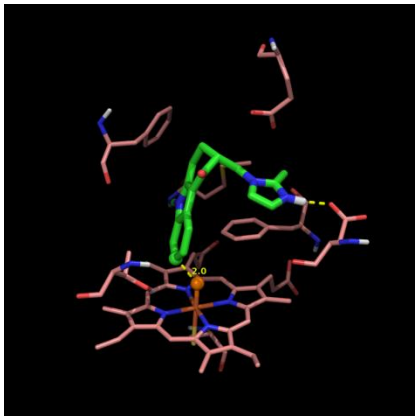
MOPPP



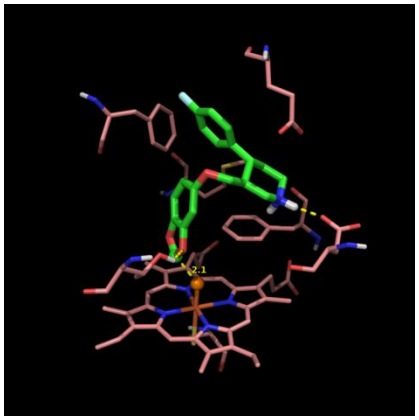
nortriptyline



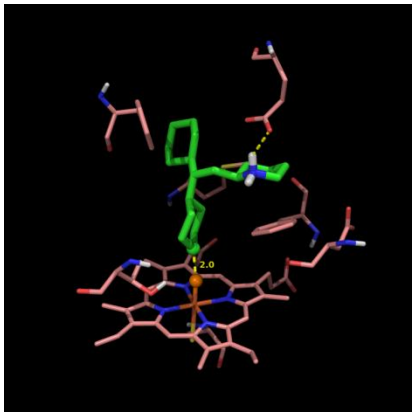
ondansetron



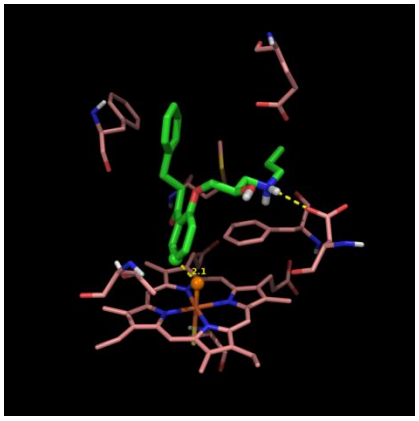
ondansetron



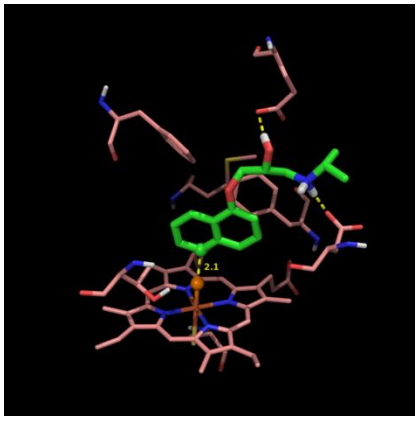
paroxetine



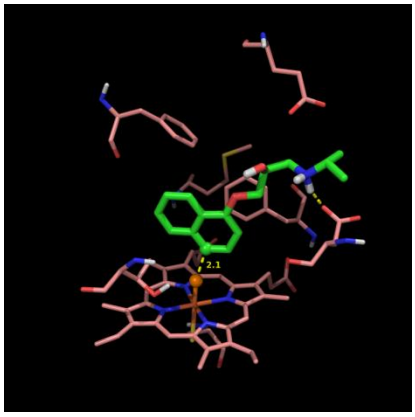
perhexiline



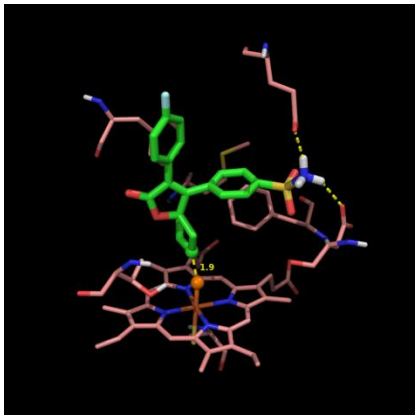
propafenone



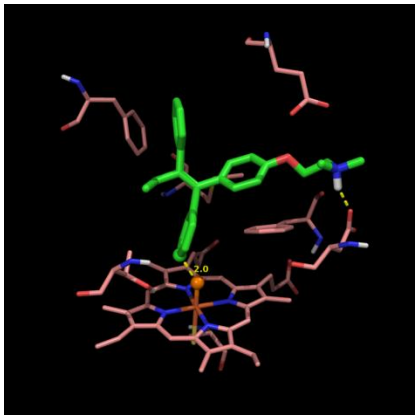
propranolol



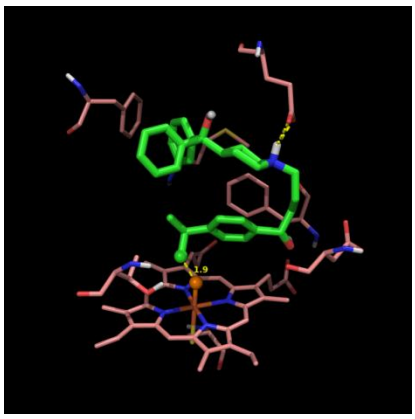
propranolol



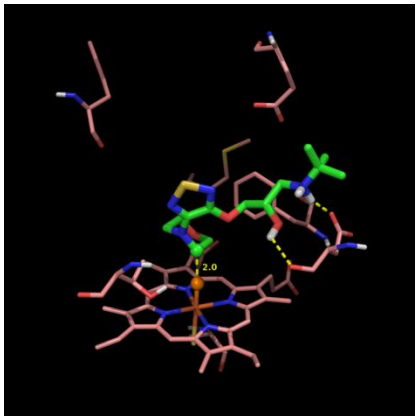
spirosulfonamide



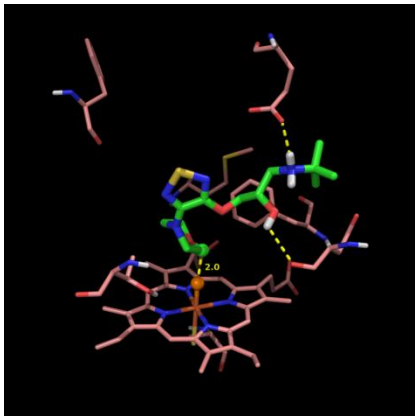
tamoxifen



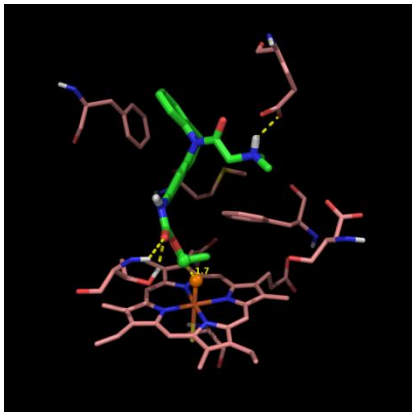
terfenadine



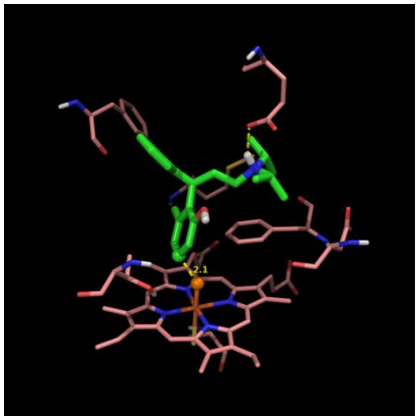
timolol



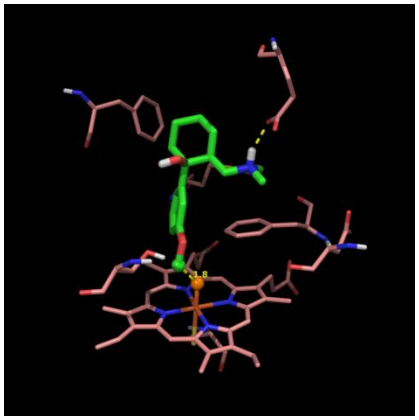
timolol



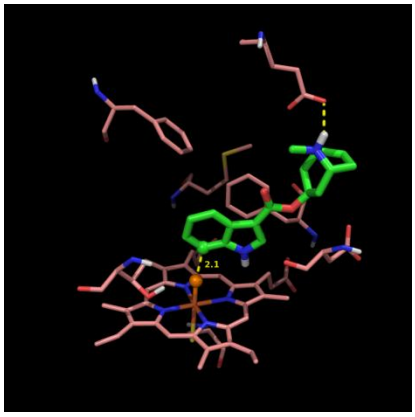
tiracizine



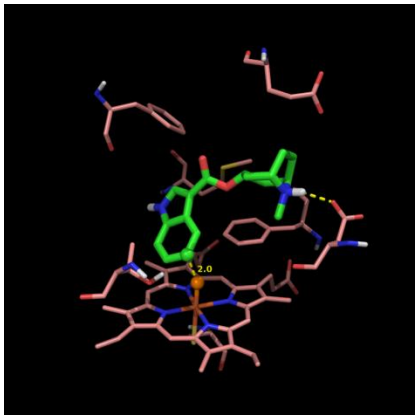
tolterodine



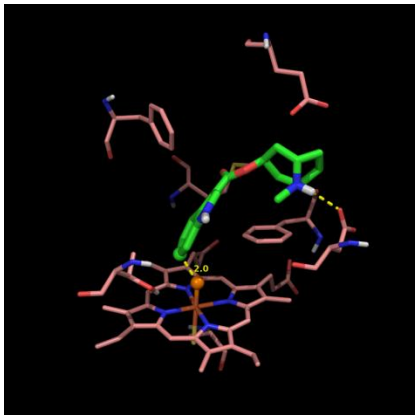
tramadol



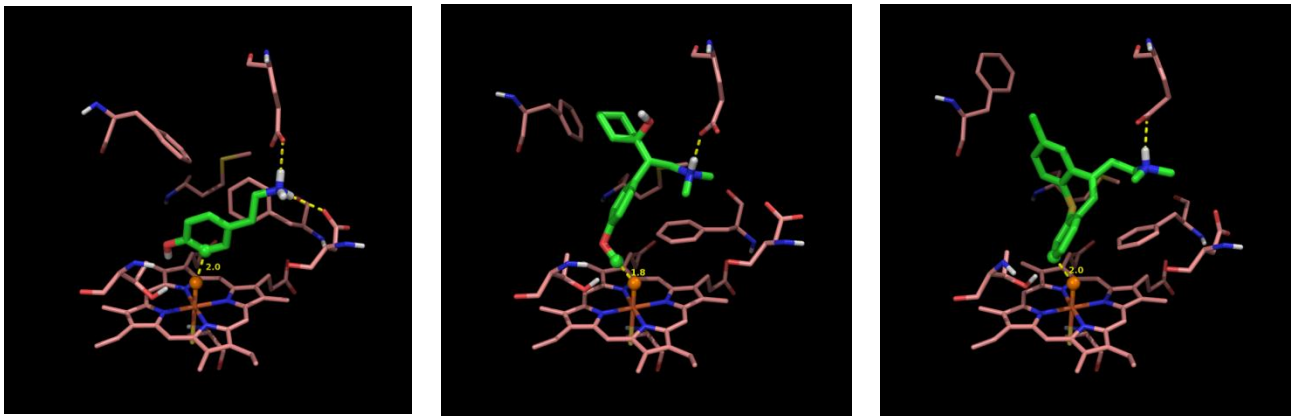
tropisetron



tropisetron



tropisetron



tyramine venlafaxine zotepine

Figure 3.21. Lowest energy poses that lead to the true positive predictions.

4. Full lists of dihedral angle changes for the cases we listed in the main text

Table 3.7. Dihedral angles of the lowest energy pose (with the SOM atom constrained) for 4-methoxyaphetamine.

4-methoxyamphetamine			Crystal Structure	Refine 1		Refine 2	
				Start	End	Start	End
119	VAL	chi1	-162.5	-162.5	-162.5	-81.8	-81.9
120	PHE	chi1	-162.6	-135.9	-138.2	-166.5	-167.1
		chi2	86.1	128.8	136.7	85.2	83.3
121	LEU	chi1	-60.2	-58.9	-58.6	-56.8	-55.7
		chi2	-179.9	-177.3	-178.6	175.9	177.6
213	LEU	chi1	-179.2	178.4	179.7	-173.1	-178.5
		chi2	60.1	56.5	56.4	55.3	53.1
214	LYS	chi1	-162.8	-179.8	-154.2	178.3	173.9
		chi2	-171.3	168.7	-171.0	-176.4	-178.0
		chi3	-168.1	160.9	90.1	144.8	154.8
		chi4	-65.7	-71.5	-168.5	-164.1	-165.8
215	GLU	chi1	-77.1	-74.6	-72.8	-78.6	-82.4
		chi2	168.7	170.4	169.7	174.6	176.5
		chi3	-16.4	-18.4	-19.8	-22.5	-15.3
216	GLU	chi1	-10.8	-68.1	-66.6	171.8	-71.8
		chi2	-171.8	-176.4	-178.9	169.2	179.9
		chi3	97.5	108.9	-68.9	-109.1	125.4
217	SER	chi1	75.9	75.9	75.9	68.1	73.4
219	PHE	chi1	-72.5	-72.5	-72.5	-76.3	-73.8
		chi2	106.0	106.0	106.0	-67.7	-59.8

220	LEU	chi1	-76.7	-76.7	-76.7	-59.1	-60.5
		chi2	164.6	164.6	164.6	-176.3	-179.6
221	ARG	chi1	63.7	63.7	63.7	-67.1	-69.5
		chi2	-169.0	-169.0	-169.0	164.2	178.9
		chi3	178.5	178.5	178.5	-148.8	-73.7
		chi4	-148.4	-148.4	-148.4	-161.1	172.1
		chi5	1.0	1.0	1.0	4.8	6.3
222	GLU	chi1	-124.0	-124.0	-124.0	-58.4	-69.2
		chi2	75.3	75.3	75.3	-58.7	-65.1
		chi3	45.3	45.3	45.3	139.9	-24.9
301	ASP	chi1	-71.6	-83.2	-79.7	-74.7	-73.9
		chi2	-170.6	179.5	-164.9	131.0	134.1
302	LEU	chi1	-70.5	-70.5	-70.5	-73.4	-72.9
		chi2	166.1	166.1	166.1	164.0	164.0
303	PHE	chi1	-99.0	-99.0	-99.0	-165.6	-165.0
		chi2	-70.6	-70.6	-70.6	99.5	98.2
304	SER	chi1	68.5	56.7	58.5	57.7	55.3
307	MET	chi1	179.2	179.2	179.2	177.0	176.7
		chi2	-165.7	-165.7	-165.7	-93.2	-93.6
		chi3	-37.7	-37.7	-37.7	-73.2	-73.9
308	VAL	chi1	63.2	63.2	63.2	63.5	69.2
309	THR	chi1	58.4	54.7	54.8	54.7	52.9
310	THR	chi1	-54.8	-54.8	-54.8	-30.7	-31.1
311	SER	chi1	177.2	177.2	177.2	-179.3	-177.5
312	THR	chi1	-65.2	-65.2	-65.2	-64.7	-67.5
313	THR	chi1	-62.4	-62.4	-62.4	-54.1	-52.3
368	ASP	chi1	-176.0	-176.0	-176.0	-61.6	-62.4
		chi2	18.7	18.7	18.7	-101.1	-100.6
369	ILE	chi1	62.1	62.1	62.1	-62.2	-50.1
		chi2	140.9	140.9	140.9	112.7	144.5
370	VAL	chi1	-134.7	-134.7	-134.7	-168.9	-164.6
371	PRO	chi1	-25.7	-25.7	-25.7	26.9	26.5
		chi2	30.6	30.6	30.6	-29.1	-28.7
372	LEU	chi1	-57.3	-57.3	-57.3	-57.0	-57.6
		chi2	-179.2	-179.2	-179.2	174.8	175.3
374	MET	chi1	179.8	179.8	179.8	176.5	177.2
		chi2	45.7	45.7	45.7	56.6	58.0
		chi3	48.9	48.9	48.9	165.3	162.7
375	THR	chi1	-174.7	-174.7	-174.7	51.8	51.9
480	VAL	chi1	-147.9	-147.9	-147.9	-70.5	-71.6

481	PHE	chi1	172.6	172.6	172.6	-85.7	-88.9
		chi2	72.7	72.7	72.7	66.2	69.8
483	PHE	chi1	-132.4	-133.2	-131.5	179.6	-179.9
		chi2	83.4	89.6	89.3	58.7	64.8
484	LEU	chi1	-156.8	-156.8	-156.8	-74.1	-83.7
		chi2	158.8	158.8	158.8	96.4	91.4
485	VAL	chi1	-30.4	-30.4	-30.4	-172.1	-172.1
486	SER	chi1	-77.2	-77.2	-77.2	-179.4	179.7

Table 3.8. Dihedral angles of the lowest energy pose (with the SOM atom constrained) for fluperlapine.

fluperlapine			Crystal Structure	Refine 1		Refine 2	
				Start	End	Start	End
120	PHE	chi1	-162.6	-140.7	-137.5	-167.5	-125.0
		chi2	86.1	142.6	137.2	117.7	159.5
215	GLU	chi1	-77.1	-77.1	-76.3	-83.4	-81.4
		chi2	168.7	172.2	172.5	175.1	179.0
		chi3	-16.4	-15.6	-17.0	-17.4	-21.3
216	GLU	chi1	-10.8	-73.3	-69.3	-71.3	-75.2
		chi2	-171.8	-166.0	-178.1	-173.0	-175.2
		chi3	97.5	-113.2	-82.5	145.7	-48.8
217	SER	chi1	75.9	75.9	75.9	75.2	-57.9
219	PHE	chi1	-72.5	-72.5	-72.5	-73.6	-62.2
		chi2	106.0	106.0	106.0	118.9	114.8
220	LEU	chi1	-76.7	-76.7	-76.7	-74.1	-54.8
		chi2	164.6	164.6	164.6	153.4	172.2
221	ARG	chi1	63.7	63.7	63.7	67.4	-147.4
		chi2	-169.0	-169.0	-169.0	-172.5	173.6
		chi3	178.5	178.5	178.5	-172.7	63.2
		chi4	-148.4	-148.4	-148.4	-135.5	-106.9
		chi5	1.0	1.0	1.0	0.5	-2.0
222	GLU	chi1	-124.0	-124.0	-124.0	177.1	-49.2
		chi2	75.3	75.3	75.3	64.8	-57.9
		chi3	45.3	45.3	45.3	8.4	144.5
301	ASP	chi1	-71.6	-79.7	-80.8	-93.0	-84.4
		chi2	-170.6	-154.7	-164.7	-34.6	-54.8
302	LEU	chi1	-70.5	-70.5	-70.5	-70.5	-71.7
		chi2	166.1	166.1	166.1	167.2	166.0
303	PHE	chi1	-99.0	-99.0	-99.0	-94.4	-163.7

		chi2	-70.6	-70.6	-70.6	-71.3	-83.8
304	SER	chi1	68.5	61.4	62.4	62.0	67.1
307	MET	chi1	179.2	179.2	179.2	177.7	-175.3
		chi2	-165.7	-165.7	-165.7	-176.8	-133.9
		chi3	-37.7	-37.7	-37.7	-48.4	67.8
308	VAL	chi1	63.2	63.2	63.2	48.9	-175.3
309	THR	chi1	58.4	50.2	50.4	44.4	46.7
310	THR	chi1	-54.8	-54.8	-54.8	-53.2	-13.9
311	SER	chi1	177.2	177.2	177.2	-179.1	51.1
312	THR	chi1	-65.2	-65.2	-65.2	-63.0	-62.6
313	THR	chi1	-62.4	-62.4	-62.4	-55.3	-56.0
368	ASP	chi1	-176.0	-176.0	-176.0	-159.0	-153.6
		chi2	18.7	18.7	18.7	-2.5	-7.4
369	ILE	chi1	62.1	62.1	62.1	78.4	-55.3
		chi2	140.9	140.9	140.9	106.4	97.0
370	VAL	chi1	-134.7	-134.7	-134.7	-156.1	-167.2
371	PRO	chi1	-25.7	-25.7	-25.7	-23.9	29.3
		chi2	30.6	30.6	30.6	27.4	-33.1
372	LEU	chi1	-57.3	-57.3	-57.3	-57.7	-60.5
		chi2	-179.2	-179.2	-179.2	177.3	174.9
374	MET	chi1	179.8	179.8	179.8	175.9	176.0
		chi2	45.7	45.7	45.7	45.0	73.0
		chi3	48.9	48.9	48.9	46.0	-84.3
375	THR	chi1	-174.7	-174.7	-174.7	-178.1	71.5
480	VAL	chi1	-147.9	-147.9	-147.9	-175.3	-176.0
481	PHE	chi1	172.6	172.6	172.6	178.9	176.9
		chi2	72.7	72.7	72.7	65.7	72.2
483	PHE	chi1	-132.4	-134.7	-134.5	-126.8	-147.0
		chi2	83.4	89.5	89.4	115.7	-10.4
484	LEU	chi1	-156.8	-156.8	-156.8	-155.1	-144.8
		chi2	158.8	158.8	158.8	164.7	141.2
485	VAL	chi1	-30.4	-30.4	-30.4	-58.5	169.3
486	SER	chi1	-77.2	-77.2	-77.2	-63.5	62.4

Table 3.9. Dihedral angles of the lowest energy pose (with the SOM atom constrained) for metoprolol (benzylic hydroxylation).

metoprolol Benzylic Hydroxylation			Crystal Structure	Refine 1		Refine 2	
				Start	End	Start	End
119	VAL	chi1	-162.5	-162.5	-162.5	-159.3	-82.2

120	PHE	chi1	-162.6	-171.2	-167.7	-178.5	-166.1
		chi2	86.1	50.0	82.0	107.4	68.6
121	LEU	chi1	-60.2	-58.3	-62.2	-69.6	-55.4
		chi2	-179.9	178.1	179.6	-172.4	178.5
213	LEU	chi1	-179.2	178.2	178.5	-172.1	179.6
		chi2	60.1	63.2	54.0	62.8	53.4
214	LYS	chi1	-162.8	177.9	176.7	177.6	173.8
		chi2	-171.3	170.6	-178.7	172.3	-176.2
		chi3	-168.1	162.8	149.7	163.4	156.7
		chi4	-65.7	-71.5	-164.8	-71.0	-167.0
215	GLU	chi1	-77.1	-76.2	-76.5	-81.8	-83.4
		chi2	168.7	169.7	171.7	174.6	175.2
		chi3	-16.4	-19.1	-19.4	-24.2	-16.0
216	GLU	chi1	-10.8	-66.7	-69.3	-68.9	-71.0
		chi2	-171.8	-170.4	-169.0	177.6	-174.4
		chi3	97.5	179.4	-14.8	120.5	-44.9
217	SER	chi1	75.9	75.9	75.9	77.7	70.0
219	PHE	chi1	-72.5	-72.5	-72.5	-68.5	-74.1
		chi2	106.0	106.0	106.0	120.4	-62.3
220	LEU	chi1	-76.7	-76.7	-76.7	-61.0	-57.1
		chi2	164.6	164.6	164.6	173.0	-175.7
221	ARG	chi1	63.7	63.7	63.7	67.0	-70.1
		chi2	-169.0	-169.0	-169.0	-172.7	174.7
		chi3	178.5	178.5	178.5	-174.0	-70.5
		chi4	-148.4	-148.4	-148.4	-138.5	-156.0
		chi5	1.0	1.0	1.0	0.3	3.1
222	GLU	chi1	-124.0	-124.0	-124.0	172.9	-68.7
		chi2	75.3	75.3	75.3	64.1	-64.9
		chi3	45.3	45.3	45.3	12.1	153.1
301	ASP	chi1	-71.6	-68.9	-91.9	-72.0	-71.0
		chi2	-170.6	-138.3	-56.2	-160.5	-56.3
302	LEU	chi1	-70.5	-70.5	-70.5	-66.8	-63.6
		chi2	166.1	166.1	166.1	168.0	168.7
303	PHE	chi1	-99.0	-99.0	-99.0	-91.9	-89.3
		chi2	-70.6	-70.6	-70.6	-69.7	111.8
304	SER	chi1	68.5	44.2	52.4	72.3	-174.5
307	MET	chi1	179.2	179.2	179.2	178.1	-166.4
		chi2	-165.7	-165.7	-165.7	-177.6	-174.2
		chi3	-37.7	-37.7	-37.7	-47.8	-45.0
308	VAL	chi1	63.2	63.2	63.2	54.7	74.1

309	THR	chi1	58.4	51.9	54.1	53.9	53.0
310	THR	chi1	-54.8	-54.8	-54.8	-46.8	-49.0
311	SER	chi1	177.2	177.2	177.2	180.0	52.1
312	THR	chi1	-65.2	-65.2	-65.2	-61.3	-66.4
313	THR	chi1	-62.4	-62.4	-62.4	-54.9	-54.0
368	ASP	chi1	-176.0	-176.0	-176.0	-161.2	-61.2
		chi2	18.7	18.7	18.7	0.6	81.0
369	ILE	chi1	62.1	62.1	62.1	80.5	-62.4
		chi2	140.9	140.9	140.9	113.7	111.1
370	VAL	chi1	-134.7	-134.7	-134.7	-77.8	-164.8
371	PRO	chi1	-25.7	-25.7	-25.7	-23.7	27.9
		chi2	30.6	30.6	30.6	27.4	-30.2
372	LEU	chi1	-57.3	-57.3	-57.3	-57.1	-56.2
		chi2	-179.2	-179.2	-179.2	178.2	176.1
374	MET	chi1	179.8	179.8	179.8	177.4	178.7
		chi2	45.7	45.7	45.7	42.8	57.5
		chi3	48.9	48.9	48.9	54.8	161.3
375	THR	chi1	-174.7	-174.7	-174.7	-178.5	52.1
480	VAL	chi1	-147.9	-147.9	-147.9	-175.0	-66.5
481	PHE	chi1	172.6	172.6	172.6	178.6	176.7
		chi2	72.7	72.7	72.7	64.9	71.5
483	PHE	chi1	-132.4	-130.2	89.3	-115.8	173.0
		chi2	83.4	89.7	-66.6	111.8	62.9
484	LEU	chi1	-156.8	-156.8	-156.8	-159.7	-84.6
		chi2	158.8	158.8	158.8	161.9	91.4
485	VAL	chi1	-30.4	-30.4	-30.4	-57.9	167.9
486	SER	chi1	-77.2	-77.2	-77.2	-61.9	55.8

Table 3.10. Dihedral angles of the lowest energy pose (with the SOM atom constrained) for metoprolol (O-demethylation).

metoprolol O-demethylation			Crystal Structure	Refine 1		Refine 2	
				Start	End	Start	End
119	VAL	chi1	-162.5	-162.5	-162.5	-162.9	-88.7
120	PHE	chi1	-162.6	-164.4	-155.1	-156.4	-162.1
		chi2	86.1	95.0	107.2	-173.7	-160.5
121	LEU	chi1	-60.2	-54.2	-51.8	-75.9	-59.0
		chi2	-179.9	177.8	174.6	-175.6	107.8
213	LEU	chi1	-179.2	175.0	177.2	-179.3	174.9
		chi2	60.1	53.7	55.2	49.4	51.6

214	LYS	chi1	-162.8	179.6	-154.3	-179.2	-161.4
		chi2	-171.3	169.5	-171.3	170.3	-169.8
		chi3	-168.1	160.9	89.6	159.9	-66.8
		chi4	-65.7	-71.6	-167.6	-71.0	174.6
215	GLU	chi1	-77.1	-74.9	-72.0	-76.3	-76.5
		chi2	168.7	170.0	170.3	173.8	174.8
		chi3	-16.4	-18.9	-20.1	-28.2	-27.0
216	GLU	chi1	-10.8	-67.6	-61.3	-72.0	-73.3
		chi2	-171.8	-170.0	-161.3	-179.5	-175.0
		chi3	97.5	59.7	62.0	125.2	148.2
217	SER	chi1	75.9	75.9	75.9	78.0	77.7
219	PHE	chi1	-72.5	-72.5	-72.5	-67.2	-67.4
		chi2	106.0	106.0	106.0	120.1	128.3
220	LEU	chi1	-76.7	-76.7	-76.7	-59.6	-64.9
		chi2	164.6	164.6	164.6	174.6	171.2
221	ARG	chi1	63.7	63.7	63.7	67.0	-66.0
		chi2	-169.0	-169.0	-169.0	-172.8	-178.6
		chi3	178.5	178.5	178.5	-174.4	-74.4
		chi4	-148.4	-148.4	-148.4	-140.1	169.6
		chi5	1.0	1.0	1.0	0.2	8.4
222	GLU	chi1	-124.0	-124.0	-124.0	172.8	175.3
		chi2	75.3	75.3	75.3	63.9	65.1
		chi3	45.3	45.3	45.3	12.6	8.2
301	ASP	chi1	-71.6	-68.4	-71.2	-77.2	-74.5
		chi2	-170.6	-158.7	-162.1	162.9	-171.7
302	LEU	chi1	-70.5	-70.5	-70.5	-70.9	-68.9
		chi2	166.1	166.1	166.1	167.0	163.6
303	PHE	chi1	-99.0	-99.0	-99.0	-91.0	-159.3
		chi2	-70.6	-70.6	-70.6	-69.1	-82.5
304	SER	chi1	68.5	63.0	54.4	58.7	59.1
307	MET	chi1	179.2	179.2	179.2	173.9	176.3
		chi2	-165.7	-165.7	-165.7	-176.7	-94.0
		chi3	-37.7	-37.7	-37.7	-51.8	-73.7
308	VAL	chi1	63.2	63.2	63.2	53.3	62.8
309	THR	chi1	58.4	44.7	53.0	48.7	51.9
310	THR	chi1	-54.8	-54.8	-54.8	-48.4	-29.6
311	SER	chi1	177.2	177.2	177.2	178.6	-179.9
312	THR	chi1	-65.2	-65.2	-65.2	-62.5	-62.4
313	THR	chi1	-62.4	-62.4	-62.4	-55.4	-53.9
368	ASP	chi1	-176.0	-176.0	-176.0	-159.1	-156.3

		chi2	18.7	18.7	18.7	-0.2	172.0
369	ILE	chi1	62.1	62.1	62.1	79.5	81.7
		chi2	140.9	140.9	140.9	107.1	105.6
370	VAL	chi1	-134.7	-134.7	-134.7	-166.6	-165.9
371	PRO	chi1	-25.7	-25.7	-25.7	-22.6	30.0
		chi2	30.6	30.6	30.6	26.4	-34.1
372	LEU	chi1	-57.3	-57.3	-57.3	-58.8	-58.3
		chi2	-179.2	-179.2	-179.2	178.7	179.0
374	MET	chi1	179.8	179.8	179.8	170.5	175.9
		chi2	45.7	45.7	45.7	47.9	51.0
		chi3	48.9	48.9	48.9	66.9	64.7
375	THR	chi1	-174.7	-174.7	-174.7	-179.0	50.4
480	VAL	chi1	-147.9	-147.9	-147.9	-174.9	-176.1
481	PHE	chi1	172.6	172.6	172.6	178.6	175.7
		chi2	72.7	72.7	72.7	64.7	78.2
483	PHE	chi1	-132.4	-134.5	78.6	-115.7	61.8
		chi2	83.4	90.4	148.4	109.7	88.1
484	LEU	chi1	-156.8	-156.8	-156.8	-158.9	156.2
		chi2	158.8	158.8	158.8	172.2	145.2
485	VAL	chi1	-30.4	-30.4	-30.4	-57.5	168.6
486	SER	chi1	-77.2	-77.2	-77.2	-60.8	179.0

Chapter 4. Concluding Remarks

This dissertation has described the development of computational approaches for practical structure-based drug discovery. The new accurate energy model (VSGB 2.0) has been presented along with advanced sampling algorithms for protein structure refinement. Given sufficient sampling, our methodology is able to select the structures with low RMSDs from the native ones. The VSGB 2.0 model has been applied to an accurate approach (IDSite) to predict P450-mediated metabolism. While competing computational methodologies for predicting SOMs are generally plagued by high false positive rates, IDSite correctly identifies almost all experimentally observed SOMs with only very few false positive predictions. Delivering such high accuracy, IDSite is likely to become practically useful in accelerating the drug discovery process. The example of IDSite shows that a highly accurate energy model in combination with efficient conformational sampling can indeed lead to transferrable, physical models for drug development. In addition to protein structure refinement and SOM prediction, the methodologies described in this dissertation can also contribute to building tools to study the mechanism of action (e.g. with QM/MM methods), to estimate binding affinities, and to assist fragment-based drug design, which will further accelerate the search of better and safer drugs.

References

(1) Jorgensen, W. L. *Science* **2004**, *303*, 1813.

(2) Prathipati, P.; Dixit, A.; Saxena, A. K. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 133.

(3) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. *Br. J. Pharmacol.* **2008**, *153*, S7.

(4) Keseru, G. M.; Makara, G. M. *Drug Discov. Today* **2006**, *11*, 741.

(5) Schneider, G.; Fechner, U. *Nat. Rev. Drug Discovery* **2005**, *4*, 649.

(6) Sun, H.; Scott, D. O. *Chem. Biol. Drug Des.* **2010**, *75*, 3.

(7) Sciabola, S.; Carosati, E.; Baroni, M.; Mannhold, R. *J. Med. Chem.* **2005**, *48*, 3756.

(8) Kumar, B.; Kotla, R.; Buddiga, R.; Roy, J.; Singh, S. S.; Gundla, R.; Ravikumar, M.; Sarma, J. *J. Mol. Model.* **2011**, *17*, 151.

(9) Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. *Curr. Med. Chem.* **2004**, *11*, 71.

(10) Gao, Q. Z.; Yang, L. L.; Zhu, Y. Q. *Curr. Comput.-Aided Drug Des.* **2010**, *6*, 37.

(11) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.

(12) Chohan, K. K.; Paine, S. W.; Waters, N. J. *Curr. Top. Med. Chem.* **2006**, *6*, 1569.

(13) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647.

(14) Klenner, A.; Hartenfeller, M.; Schneider, P.; Schneider, G. *Drug Discovery Today: Technologies* **2010**, *7*, e237.

(15) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. *Acc. Chem. Res.* **1994**, *27*, 117.

(16) Verlinde, C.; Hol, W. G. J. *Structure* **1994**, *2*, 577.

(17) Klebe, G. *J. Mol. Med.* **2000**, *78*, 269.

(18) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739.

(19) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.* **2004**, *47*, 1750.

(20) Goodsell, D. S.; Morris, G. M.; Olson, A. J. *J. Mol. Recognit.* **1996**, *9*, 1.

(21) Good, A. C.; Krystek, S. R.; Mason, J. S. *Drug Discov. Today* **2000**, *5*, S61.

(22) Tang, Y. T.; Marshall, G. R. *J. Chem. Inf. Model.* **2011**, *51*, 214.

(23) Sander, C.; Schneider, R. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 56.

(24) Polyak, K.; Xia, Y.; Zweier, J. L.; Kinzler, K. W.; Vogelstein, B. *Nature* **1997**, *389*, 300.

(25) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. *Nucleic Acids Res.* **2003**, *31*, 3381.

(26) Jayatilleke, P. R. N.; Nair, A. C.; Welsh, W. J. *Abstr. Pap. Am. Chem. Soc.* **2000**, *219*, 129.

(27) Tomasselli, A. G.; Heinrikson, R. L. *Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology* **2000**, *1477*, 189.

(28) Liang, J.; Edelsbrunner, H.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 18.

(29) Movshovitz-Attias, D.; London, N.; Schueler-Furman, O. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1939.

(30) Kidd, B. A.; Baker, D.; Thomas, W. E. *PLoS Comput. Biol.* **2009**, *5*.

(31) Spaar, A.; Helms, V. *J. Chem. Theory Comput.* **2005**, *1*, 723.

(32) Zhou, W. Q.; Yan, H. *Bioinformatics* **2010**, *26*, 2541.

(33) Sadowski, M. I.; Jones, D. T. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 476.

(34) Cheng, A.; Diller, D. J.; Dixon, S. L.; Egan, W. J.; Lauri, G.; Merz, K. M. *J. Comput. Chem.* **2002**, *23*, 172.

(35) Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. *Mol. Pharm.* **2008**, *5*, 266.

(36) Ioannides, C. In *Chemistry and Molecular Aspects of Drug Design and Action*; CRC Press: 2008, p 253.

(37) Wang, B.; Yang, L.-P.; Zhang, X.-Z.; Huang, S.-Q.; Bartlam, M.; Zhou, S.-F. *Drug Metab. Rev.* **2009**, *41*, 573.

(38) Vermeulen, N. P. E. *Curr. Top. Med. Chem.* **2003**, *3*, 1227.

(39) Zamora, I.; Afzelius, L.; Cruciani, G. *J. Med. Chem.* **2003**, *46*, 2313.

(40) de Groot, M. J. *Drug Discov. Today* **2006**, *11*, 601.

(41) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. *J. Med. Chem.* **2007**, *50*, 3173.

(42) Bochevarov, A. D.; Li, J. N.; Song, W. J.; Friesner, R. A.; Lippard, S. J. *J. Am. Chem. Soc.* **2011**, *133*, 7384.

(43) Patard, L.; Stoven, V.; Gharib, B.; Bontems, F.; Lallemand, J. Y.; DeReggi, M. *Protein Eng.* **1996**, *9*, 949.

(44) Evers, A.; Klabunde, T. *J. Med. Chem.* **2005**, *48*, 1088.

(45) Muegge, I.; Enyedy, I. J. *Curr. Med. Chem.* **2004**, *11*, 693.

(46) Blundell, T. L.; Jhoti, H.; Abell, C. *Nat. Rev. Drug Discov.* **2002**, *1*, 45.

(47) Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.; Olde, B.; Moreno, R. F.; Kerlavage, A. R.; McCombie, W. R.; Venter, J. C. *Science* **1991**, *252*, 1651.

(48) Sander, C.; Schneider, R. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 56.

(49) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779.

(50) Rayan, A.; Noy, E.; Chema, D.; Levitzki, A.; Goldblum, A. *Curr. Med. Chem.* **2004**, *11*, 675.

(51) Rockey, W. M.; Elcock, A. H. *Curr. Protein Pept. Sci.* **2006**, *7*, 437.

(52) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 15.

(53) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(54) Zhou, R. H. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 148.

(55) Zhou, R. H.; Berne, B. J. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12777.

(56) Huang, A.; Stultz, C. M. *Biophys. J.* **2007**, *92*, 34.

(57) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(58) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005.

(59) Tannor, D. J.; Marten, B.; Murphy, R.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Ringnalda, M.; Goddard, W. A.; Honig, B. *J. Am. Chem. Soc.* **1994**, *116*, 11875.

(60) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983.

(61) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479.

(62) Gallicchio, E.; Paris, K.; Levy, R. M. *J. Chem. Theory Comput.* **2009**, *5*, 2544.

(63) Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. *J. Phys. Chem. B* **2002**, *106*, 11673.

(64) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 351.

(65) Yu, Z. Y.; Jacobson, M. P.; Friesner, R. A. *J. Comput. Chem.* **2006**, *27*, 72.

(66) Zhu, K.; Pincus, D. L.; Zhao, S. W.; Friesner, R. A. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 438.

(67) Li, X.; Jacobson, M. P.; Zhu, K.; Zhao, S. W.; Friesner, R. A. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 824.

(68) Zhu, K.; Shirts, M. R.; Friesner, R. A. *J. Chem. Theory Comput.* **2007**, *3*, 2108.

(69) Felts, A. K.; Gallicchio, E.; Chekmarev, D.; Paris, K. A.; Friesner, R. A.; Levy, R. M. *J. Chem. Theory Comput.* **2008**, *4*, 855.

(70) Sellers, B. D.; Zhu, K.; Zhao, S.; Friesner, R. A.; Jacobson, M. P. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 959.

(71) Fiser, A.; Do, R. K. G.; Sali, A. *Protein Sci.* **2000**, *9*, 1753.

(72) Galaktionov, S.; Nikiforovich, G. V.; Marshall, G. R. *Peptide Science* **2001**, *60*, 153.

(73) Im, W. P.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691.

(74) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265.

(75) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6760.

(76) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 115.

(77) Zhang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591.

(78) Jacobson, M. P.; Friesner, R. A.; Xiang, Z. X.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597.

(79) Xiang, Z. X.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421.

(80) Zhao, S. W.; Zhu, K.; Li, J. N.; Friesner, R. A. *Proteins: Struct., Funct., Bioinf.* **2011**, (in press).

(81) Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. In *Numerical Computer Methods, Pt D* 2004; Vol. 383, p 66.

(82) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *The Journal of Physical Chemistry B* **2001**, *105*, 6474.

(83) Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1994**, *3*, 211.

(84) Paton, R. S.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944.

(85) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549.

(86) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 6946.

(87) Morozov, A. V.; Kortemme, T. *Adv. Protein Chem.* **2005**, *72*, 1.

(88) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. *Nucleic Acids Res.* **2006**, *34*, 564.

(89) Churchill, C. D. M.; Navarro-Whyte, L.; Rutledge, L. R.; Wetmore, S. D. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10657.

(90) Wang, L. J.; Sun, N.; Terzyan, S.; Zhang, X. J.; Benson, D. R. *Biochemistry* **2006**, *45*, 13750.

(91) Gazit, E. *FASEB J.* **2002**, *16*, 77.

(92) Magalhaes, A.; Maigret, B.; Hoflack, J.; Gomes, J. N. F.; Scheraga, H. A. *J. Protein Chem.* **1994**, *13*, 195.

(93) Sinnokrot, M. O.; Sherrill, C. D. *The Journal of Physical Chemistry A* **2004**, *108*, 10200.

(94) Kawakami, J.; Okabe, S.; Tanabe, Y.; Sugimoto, N. *Nucleosides Nucleotides Nucleic Acids* **2008**, *27*, 292.

(95) Chelli, R.; Gervasio, F. L.; Procacci, P.; Schettino, V. *J. Am. Chem. Soc.* **2002**, *124*, 6133.

(96) Burley, S. K.; Petsko, G. A. *Science* **1985**, *229*, 23.

(97) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

(98) Pal, T. K.; Sankararamakrishnan, R. *J. Mol. Graph. Model.* **2008**, *27*, 20.

(99) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609.

(100) Qin, B. Y.; Bewley, M. C.; Creamer, L. K.; Baker, H. M.; Baker, E. N.; Jameson, G. B. *Biochemistry* **1998**, *37*, 14014.

(101) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167.

(102) Onuchic, J. N.; LutheySchulten, Z.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.

(103) Bailey, D. G.; Malcolm, J.; Arnold, O.; Spence, J. D. *Br. J. Clin. Pharmacol.* **1998**, *46*, 101.

(104) Preskorn, S. H. *Clin. Pharmacokinet.* **1997**, *32*, 1.

(105) Dresser, G. K.; Spence, J. D.; Bailey, D. G. *Clin. Pharmacokinet.* **2000**, *38*, 41.

(106) Afzelius, L.; Arnby, C. H.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Raubacher, F.; Weidolf, L. *Drug Metab. Rev.* **2007**, *39*, 61.

(107) Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. *J. Med. Chem.* **2003**, *46*, 1330.

(108) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. *ACS Medicinal Chemistry Letters* **2010**, *1*, 96.

(109) de Groot, M. J.; Alex, A. A.; Jones, B. C. *J. Med. Chem.* **2002**, *45*, 1983.

(110) Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-w.; Bennett, K. P.; Breneman, C. M. *J. Chem. Inf. Model.* **2011**, *51*, 1667.

(111) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. *J. Med. Chem.* **2005**, *48*, 6970.

(112) Jones, J. P.; Korzekwa, K. R. In *Methods Enzymol.*; Eric, F. J., Michael, R. W., Eds.; Academic Press: New York, NY, 1996; Vol. Volume 272, p 326.

(113) Olâh, J.; Mulholland, A. J.; Harvey, J. N. *Proceedings of the National Academy of Sciences* **2011**, *108*, 6050.

(114) Kirton, S. B.; Kemp, C. A.; Tomkinson, N. P.; St.-Gallay, S.; Sutcliffe, M. J. *Proteins: Struct., Funct., Bioinf.* **2002**, *49*, 216.

(115) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. *J. Med. Chem.* **2006**, *49*, 2417.

(116) Unwalla, R.; Cross, J.; Salaniwal, S.; Shilling, A.; Leung, L.; Kao, J.; Humblet, C. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 237.

(117) Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Jorgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C. *J. Chem. Inf. Model.* **2009**, *49*, 43.

(118) Rydberg, P.; Hansen, S. M.; Kongsted, J.; Norrby, P. O.; Olsen, L.; Ryde, U. *J. Chem. Theory Comput.* **2008**, *4*, 673.

(119) Gleeson, M. P.; Davis, A. M.; Chohan, K. K.; Paine, S. W.; Boyer, S.; Gavaghan, C. L.; Arnby, C. H.; Kankkonen, C.; Albertson, N. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 559.

(120) Li, J.; Abel, R.; Zhu, K.; Cao, Y.; Friesner, R. *Proteins: Struct., Funct., Bioinf.* **2011**, (in press).

(121) Bathelt, C. M.; Mulholland, A. J.; Harvey, J. N. *J. Phys. Chem. A* **2008**, *112*, 13149.

(122) Tian, L.; Friesner, R. A. *J. Chem. Theory Comput.* **2009**, *5*, 1421.

(123) Glide, version 5.6; Schrödinger, Inc.: New York, NY, 2010.

(124) Prime, version 3.0; Schrödinger, Inc. : New York, NY, 2011.

(125) Paine, M. J. I.; McLaughlin, L. A.; Flanagan, J. U.; Kemp, C. A.; Sutcliffe, M. J.; Roberts, G. C. K.; Wolf, C. R. *J. Biol. Chem.* **2003**, *278*, 4021.

(126) Jaguar, version 7.6; Schrödinger, Inc.: New York, NY, 2010.

(127) Duane, S.; Kennedy, A. D.; Pendleton, B. J.; Roweth, D. *Phys. Lett. B* **1987**, *195*, 216.

(128) Guengerich, F. P.; Miller, G. P.; Hanna, I. H.; Martin, M. V.; Leger, S.; Black, C.; Chauret, N.; Silva, J. M.; Trimble, L. A.; Yergey, J. A.; Nicoll-Griffith, D. A. *Biochemistry* **2002**, *41*, 11025.

(129) Guengerich, F. P.; Hanna, I. H.; Martin, M. V.; Gillam, E. M. J. *Biochemistry* **2003**, *42*, 1245.

(130) Guengerich, F. P. *Chem. Res. Toxicol.* **2001**, *14*, 611.

(131) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279.

(132) Wang, Y. H.; Li, Y.; Wang, B. *J. Phys. Chem. B* **2007**, *111*, 4251.

(133) Schneebeli, S. T.; Hall, M. L.; Breslow, R.; Friesner, R. *J. Am. Chem. Soc.* **2009**, *131*, 3965.

(134) Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U. *J. Med. Chem.* **2006**, *49*, 6489.

(135) Rydberg, P.; Ryde, U.; Olsen, L. *J. Phys. Chem. A* **2008**, *112*, 13058.

(136) Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M. *J. Biol. Chem.* **2006**, *281*, 7614.

(137) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. *J. Med. Chem.* **1999**, *42*, 1515.

(138) Feifel, N.; Kucher, K.; Fuchs, L.; Jedrychowski, M.; Schmidt, E.; Antonin, K. H.; Bieck, P. R.; Gleiter, C. H. *Eur. J. Clin. Pharmacol.* **1993**, *45*, 265.

(139) Olesen, O. V.; Linnet, K. *Drug Metab. Dispos.* **1997**, *25*, 740.

(140) Geertsen, S.; Foster, B. C.; Wilson, D. L.; Cyr, T. D.; Casley, W. *Xenobiotica* **1995**, *25*, 895.

(141) Wolff, T.; Distlerath, L. M.; Worthington, M. T.; Guengerich, F. P. *Arch. Toxicol.* **1987**, *60*, 89.

(142) Jack, D. B.; Stenlake, J. B.; Templeto.R *Xenobiotica* **1972**, *2*, 35.

(143) Eichelbaum, M. *Fed. Proc.* **1984**, *43*, 2298.

(144) Wu, D.; Otton, S. V.; Inaba, T.; Kalow, W.; Sellers, E. M. *Biochem. Pharmacol.* **1997**, *53*, 1605.

(145) Coutts, R. T.; Bach, M. V.; Baker, G. B. *Xenobiotica* **1997**, *27*, 33.

(146) Ebner, T.; Eichelbaum, M. *Br. J. Clin. Pharmacol.* **1993**, *35*, 426.

(147) Ring, B. J.; Gillespie, J. S.; Eckstein, J. A.; Wrighton, S. A. *Drug Metab. Dispos.* **2002**, *30*, 319.

(148) Erickson, D. A.; Hollfelder, S.; Tenge, J.; Gohdes, M.; Burkhardt, J. J.; Krieter, P. A. *Drug Metab. Dispos.* **2007**, *35*, 2232.

(149) Mautz, D. S.; Nelson, W. L.; Shen, D. D. *Drug Metab. Dispos.* **1995**, *23*, 513.

(150) Yamazaki, H.; Guo, Z.; Persmark, M.; Mimura, M.; Inoue, K.; Guengerich, F. P.; Shimada, T. *Mol. Pharmacol.* **1994**, *46*, 568.

(151) Appanna, G.; Tang, B. K.; Muller, R.; Kalow, W. *Drug Metab. Dispos.* **1996**, *24*, 303.

(152) Kudo, S.; Uchida, M.; Odomi, M. *Eur. J. Clin. Pharmacol.* **1997**, *52*, 479.

(153) Oldham, H. G.; Clarke, S. E. *Drug Metab. Dispos.* **1997**, *25*, 970.

(154) Yoshii, K.; Kobayashi, K.; Tsumuji, M.; Tani, M.; Shimada, N.; Chiba, K. *Life Sci.* **2000**, *67*, 175.

(155) Kariya, S.; Isozaki, S.; Uchino, K.; Suzuki, T.; Narimatsu, S. *Biol. Pharm. Bull.* **1996**, *19*, 1511.

(156) Nielsen, K. K.; Flinois, J. P.; Beaune, P.; Brøsen, K. *J. Pharmacol. Exp. Ther.* **1996**, *277*, 1659.

(157) Gasche, Y.; Daali, Y.; Fathi, M.; Chiappe, A.; Cottini, S.; Dayer, P.; Desmeules, J. *N. Engl. J. Med.* **2004**, *351*, 2827.

(158) Su, P.; Baker, G. B.; Daneshtalab, M. *Xenobiotica* **1993**, *23*, 1289.

(159) Schmider, J.; Greenblatt, D. J.; Fogelman, S. M.; Von Moltke, L. L.; Shader, R. I. *Biopharm. Drug Dispos.* **1997**, *18*, 227.

(160) Kirkwood, L. C.; Nation, R. L.; Somogyi, A. A. *Br. J. Clin. Pharmacol.* **1997**, *44*, 549.

(161) Keizers, P. H. J.; Van Dijk, B. R.; De Graaf, C.; Van Vugt-Lussenburg, B. M. A.; Vermeulen, N. P. E.; Commandeur, J. N. M. *Xenobiotica* **2006**, *36*, 763.

(162) Funck-Brentano, C.; Turgeon, J.; Woosley, R. L.; Roden, D. M. *J. Pharmacol. Exp. Ther.* **1989**, *249*, 134.

(163) Liu, Z. R.; Mortimer, O.; Smith, C. A. D.; Wolf, C. R.; Rane, A. *Br. J. Clin. Pharmacol.* **1995**, *39*, 77.

(164) Fischer, V.; Vogels, B.; Maurer, G.; Tynes, R. E. *J. Pharmacol. Exp. Ther.* **1992**, *260*, 1355.

(165) Yu, A.-M.; Idle, J. R.; Krausz, K. W.; Küpfer, A.; Gonzalez, F. J. *J. Pharmacol. Exp. Ther.* **2003**, *305*, 315.

(166) Kaplan, H. L.; Busto, U. E.; Baylon, G. J.; Cheung, S. W.; Otton, S. V.; Somer, G.; Sellers, E. M. *J. Pharmacol. Exp. Ther.* **1997**, *281*, 103.

(167) Hutchinson, M. R.; Menelaou, A.; Foster, D. J. R.; Coller, J. K.; Somogyi, A. A. *Br. J. Clin. Pharmacol.* **2004**, *57*, 287.

(168) Obach, R. S.; Pablo, J.; Mash, D. C. *Drug Metab. Dispos.* **1998**, *26*, 764.

(169) Pierce, D.; Smith, S.; Franklin, R. *Eur. J. Clin. Pharmacol.* **1987**, *33*, 59.

(170) Colado, M. I.; Williams, J. L.; Green, A. R. *Br. J. Pharmacol.* **1995**, *115*, 1281.

(171) Lin, L. Y.; Di Stefano, E. W.; Schmitz, D. A.; Hsu, L.; Ellis, S. W.; Lennard, M. S.; Tucker, G. T.; Cho, A. K. *Drug Metab. Dispos.* **1997**, *25*, 1059.

(172) Broly, F.; Libersa, C.; Lhermitte, M.; Dupuis, B. *Biochem. Pharmacol.* **1990**, *39*, 1045.

(173) Nakajima, M.; Kobayashi, K.; Shimada, N.; Tokudome, S.; Yamamoto, T.; Kuroiwa, Y. *Br. J. Clin. Pharmacol.* **1998**, *46*, 55.

(174) Koyama, E.; Chiba, K.; Tani, M.; Ishizaki, T. *J. Pharmacol. Exp. Ther.* **1996**, *278*, 21.

(175) Dahl, M.-L.; Voortman, G.; Alm, C.; Elwin, C.-E.; Delbressine, L.; Vos, R.; Bogaards, J. J. P.; Bertilsson, L. *Clin. Drug Investig.* **1997**, *13*, 37.

(176) Springer, D.; Staack, R. F.; Paul, L. D.; Kraemer, T.; Maurer, H. H. *Xenobiotica* **2003**, *33*, 989.

(177) Fischer, V.; Vickers, A. E.; Heitz, F.; Mahadevan, S.; Baldeck, J. P.; Minery, P.; Tynes, R. *Drug Metab. Dispos.* **1994**, *22*, 269.

(178) Lalovic, B.; Phillips, B.; Risler, L. L.; Howald, W.; Shen, D. D. *Drug Metab. Dispos.* **2004**, *32*, 447.

(179) Sindrup, S. H.; Brosen, K.; Gram, L. F.; Hallas, J.; Skjelbo, E.; Allen, A.; Allen, G. D.; Cooper, S. M.; Mellows, G.; Tasker, T. C. G.; Zussman, B. D. *Clin. Pharmacol. Ther.* **1992**, *51*, 278.

(180) Amoah, A. G. B.; Gould, B. J.; Parke, D. V.; Lockhart, J. D. F. *Xenobiotica* **1986**, *16*, 63.

(181) Botsch, S.; Gautier, J. C.; Beaune, P.; Eichelbaum, M.; Kroemer, H. K. *Mol. Pharmacol.* **1993**, *43*, 120.

(182) Masubuchi, Y.; Hosokawa, S.; Horie, T.; Suzuki, T.; Ohmori, S.; Kitada, M.; Narimatsu, S. *Drug Metab. Dispos.* **1994**, *22*, 909.

(183) Crewe, H. K.; Ellis, S. W.; Lennard, M. S.; Tucker, G. T. *Biochem. Pharmacol.* **1997**, *53*, 171.

(184) Jones, B. C.; Hyland, R.; Ackland, M.; Tyman, C. A.; Smith, D. A. *Drug Metab. Dispos.* **1998**, *26*, 875.

(185) Volotinen, M.; Turpeinen, M.; Tolonen, A.; Uusitalo, J.; Mäenpää, J.; Pelkonen, O. *Drug Metab. Dispos.* **2007**, *35*, 1135.

(186) Berndt, A.; Hoffmann, C.; Richter, K.; Oertel, R.; Vierkant, A.; Siegmund, W. *Br. J. Clin. Pharmacol.* **1995**, *40*, 287.

(187) Postlind, H.; Danielson, Å.; Lindgren, A.; Andersson, S. H. G. *Drug Metab. Dispos.* **1998**, *26*, 289.

(188) Allegaert, K.; Anderson, B. J.; Verbesselt, R.; Debeer, A.; de Hoon, J.; Devlieger, H.; Van Den Anker, J. N.; Tibboel, D. *Br. J. Anaesth.* **2005**, *95*, 231.

(189) Hiroi, T.; Imaoka, S.; Funae, Y. *Biochem. Biophys. Res. Commun.* **1998**, *249*, 838.

(190) Ball, S. E.; Ahern, D.; Scatina, J.; Kao, J. *Br. J. Clin. Pharmacol.* **1997**, *43*, 619.

(191) Fogelman, S. M.; Schmider, J.; Venkatakrishnan, K.; von Moltke, L. L.; Harmatz, J. S.; Shader, R. I.; Greenblatt, D. J. *Neuropsychopharmacology* **1999**, *20*, 480.

(192) Shiraga, T.; Kaneko, H.; Iwasaki, K.; Tozuka, Z.; Suzuki, A.; Hata, T. *Xenobiotica* **1999**, *29*, 217.